

**ROBUST METHODS FOR SENSING AND
RECONSTRUCTING SPARSE SIGNALS**

by

Rafael E. Carrillo

A dissertation submitted to the Faculty of the University of Delaware in
partial fulfillment of the requirements for the degree of Doctor of Philosophy in
Electrical and Computer Engineering

Fall 2011

© 2011 Rafael E. Carrillo
All Rights Reserved

**ROBUST METHODS FOR SENSING AND
RECONSTRUCTING SPARSE SIGNALS**

by

Rafael E. Carrillo

Approved: _____
Kenneth E. Barner, Ph.D.
Chair of the Department of Electrical and Computer Engineering

Approved: _____
Babatunde Ogunnaike, Ph.D.
Interim Dean of the College of Engineering

Approved: _____
Charles G. Riordan, Ph.D.
Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Kenneth E. Barner, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Gonzalo Arce, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Javier Garcia-Frias, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Tuncer Can Aysal, Ph.D.
Member of dissertation committee

ACKNOWLEDGEMENTS

Many people has contributed to the completion of this dissertation and I am grateful to all of them. First and foremost, I would like to thank my advisor, Kenneth Barner, for providing the encouragement, supervision and support needed during my Ph.D. studies. His optimistic look into life and positive thinking always inspired me. He gave me the freedom to pick the topic I wanted and always pushed me to think outside the box. I would like to thank my committee members, Gonzalo Arce, Javier Garcia-Frias and Can Aysal for dedicating their time to read my thesis and providing useful perspectives and comments to the work presented to them. A special thanks goes to Can Aysal for helping me out along the way with many obstacles I faced.

I also want to thank the people from my lab for providing such a great working environment: Luisa Polania, Jinglun Gao, Yin Zhou, Rui Hu and Kai Liu. Many friends at Delaware made my life in Newark a delightful experience: Claudia, Andres, Mileva, Rodrigo, Alejandra, Alejandrina, Inaki, Luisa, Paola, Daniel, Fernando, Elkin, Anita, Andrea, Ivan, Gonchi, Mercedes, Melissa, Alvaro, Diego, Felipe, Cesar and many more. *Gracias a todos.*

Last but not least, I thank God and my family for always being there with me through every step of this journey. I am so thankful to have such a caring Mom and Dad who although are far away have always expressed their love and encouragement to push me to work harder. I wouldn't be what I am today if it weren't for them. I am also deeply grateful to my amazing wife Viviana for her unending love, support, and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES	x
LIST OF TABLES	xvi
ABSTRACT	xvii

Chapter

1 INTRODUCTION	1
1.1 Motivation	3
1.2 Dissertation Objectives and Contributions	5
1.3 Organization	7
2 COMPRESSED SENSING BACKGROUND	10
2.1 Introduction	10
2.2 The Sensing and Reconstruction Problems	12
2.3 Incoherence and Sampling of Sparse Signals	14
2.3.1 Sparsity	15
2.3.2 Incoherent sampling	17
2.4 Reconstruction Methods	20
2.4.1 Notation	20
2.4.2 Reconstruction in the Noiseless Case	20
2.4.3 Reconstruction from Noisy Measurements	24
2.5 Connections of CS with Other Fields	31
2.6 Concluding Remarks	34

3	A GENERALIZED CAUCHY DISTRIBUTION FRAMEWORK FOR PROBLEMS REQUIRING ROBUST BEHAVIOR	36
3.1	Introduction	36
3.2	Distributions, Optimal Filtering and M-Estimation	40
3.2.1	M-Estimation	40
3.2.2	Generalized Gaussian Distribution	41
3.2.3	Generalized Cauchy Distribution	43
3.2.4	Statistical Relationship Between the Generalized Cauchy and Gaussian Distributions	45
3.3	Generalized Cauchy Based Robust Estimation and Filtering	45
3.3.1	Generalized Cauchy Based M-Estimation	45
3.3.2	Robustness and Analysis of M-GC Estimators	50
3.3.3	Weighted M-GC Estimators	54
3.3.4	Multi-parameter Estimation	55
3.4	Robust Distance Metrics	59
3.5	Illustrative Application Areas	63
3.5.1	Robust Filtering	64
3.5.2	Robust Blind Decentralized Estimation	65
3.5.3	Robust Clustering	71
3.6	Concluding Remarks	74
4	ROBUST SAMPLING AND RECONSTRUCTION METHODS FOR SPARSE SIGNALS IN THE PRESENCE OF IMPULSIVE NOISE	76
4.1	Introduction	76
4.2	Background and Motivation	80
4.2.1	Compressed Sensing Review	80

4.2.2	Impulsive Noise in CS	83
4.3	Robust Sampling Functions	87
4.3.1	Myriad Projections	89
4.3.2	Asymptotical analysis and parameter tuning	90
4.4	Robust Reconstruction Algorithms	93
4.4.1	Lorentzian constrained L_1 minimization	94
4.4.2	Analysis under the Cauchy model	96
4.4.3	Debiasing	99
4.5	Experimental Results	100
4.5.1	Robust Sampling: Myriad Measurements	101
4.5.2	Robust Reconstruction: Lorentzian BP	112
4.6	Concluding Remarks	119
5	ROBUST BAYESIAN COMPRESSED SENSING USING GENERALIZED CAUCHY MODELS	121
5.1	Introduction	121
5.2	Bayesian Modeling and Compressed Sensing	124
5.3	Bayesian Compressed Sensing with Generalized Cauchy Priors	125
5.3.1	MAP estimation with generalized Cauchy priors	125
5.3.2	Algorithm formulation	126
5.4	Robust Bayesian Compressed Sensing with Generalized Cauchy Models	130
5.4.1	MAP estimation with generalized Cauchy priors and noise models	130
5.4.2	Fixed point algorithm	131
5.5	Experimental Results	133
5.5.1	Noiseless and light-tailed noise cases	133

5.5.2	Heavy-tailed noise	140
5.6	Concluding Remarks	142
6	LORENTZIAN ITERATIVE HARD THRESHOLDING: ROBUST COMPRESSED SENSING WITH PRIOR INFORMATION . . .	144
6.1	Introduction	144
6.2	Background and Motivation	147
6.2.1	Lorentzian Based Basis Pursuit	147
6.2.2	Iterative hard thresholding	149
6.2.3	Compressed sensing with partially known support	149
6.3	Lorentzian based Iterative Hard Thresholding Algorithm	151
6.3.1	Algorithm formulation and stability guarantees	151
6.3.2	Parameter tuning	154
6.4	Lorentzian Iterative Hard Thresholding with Prior Information	157
6.4.1	Lorentzian iterative hard thresholding with partially known support	157
6.4.2	Extension of Lorentzian iterative hard thresholding to model-sparse signals	159
6.5	Experimental Results	159
6.5.1	Robust Reconstruction: LIHT	159
6.5.2	LIHT with Partially Known Support	164
6.6	Concluding Remarks	170
7	CONCLUSIONS AND FUTURE WORK	173
7.1	Conclusions	173
7.2	Future Work	176
	BIBLIOGRAPHY	178

APPENDICES	193
A PROOF OF LEMMA 1: STATISTICAL RELATION BETWEEN GGD AND GCD RANDOM VARIABLES	194
B PROOF OF PROPOSITION 1: PROPERTIES OF THE M-GC COST FUNCTION	196
C ITERATIVE ALGORITHMS FOR CS WITH PARTIALLY KNOWN SUPPORT	199
C.1 OMP	199
C.2 CoSaMP	200
C.3 RWLS- SL_0	201
D PROOF OF THEOREM 8: STABILITY OF THE LIHT-PKS ALGORITHM	203

LIST OF FIGURES

1.1	Example of sparse signal in the frequency domain. Spectrum magnitude of $x(t)$ for $f_1 = 1/8$ and $f_2 = 1/4$	2
1.2	Block diagram of noise contributions in compressed sensing.	4
2.1	Image sparsity example. (a) Original 512×512 image with 8 bits per pixel. (b) Wavelet transform coefficients. (c) Ordered wavelet transform coefficients (absolute value) in logarithmic scale. (d) Reconstructed image by zeroing all but the 12.500 largest coefficients.	16
3.1	Typical M-GC objective functions for different values of $p \in \{0.5, 1, 1.5, 2\}$ (from bottom to top respectively). Input samples are $\mathbf{x} = [4.9, 0, 6.5, 10.0, 9.5, 1.7, 1]$ and $\sigma = 1$	47
3.2	Influence functions of the M-GC estimator for different values of p . (black:) $p = 0.5$, (blue:) $p = 1$, (red:) $p = 1.5$, and (cyan:) $p = 2$	51
3.3	Multi-parameter estimation MSE iteration evolution for a GCD process with $(\theta, \sigma, p) = (0, 1, 1.5)$	60
3.4	Contour plots of different metrics for two dimensions: (a) L_2 , (b) LL_2 (Lorentzian), (c) L_1 , and (d) LL_1 norms.	63
3.5	Power line communication enhancement. MSE for different filtering structures as function of the tail parameter α	65
3.6	Power line communication enhancement. (a) Transmitted signal, (b) Received signal corrupted by α -stable noise $\alpha = 0.4$ Filtering results with: (c) Mean, (d) Median, (e) FLOM $p = 0.25$ (f) Myriad, (g) Meridian, (h) M-GC.	66

3.7	Sensor network example with parameters: $\theta = 1$, $\tau = 0$, $\sigma_n = 1$ and $K = 1000$. Comparison of MLUGC, MLUG, BE and CE. (a) Channel noise contaminated p -Gaussian distributed with $\sigma_w^2 = 0.5$. MSE as function of the of the contamination parameter, p . (b) Channel noise α -stable distributed with $\sigma_w = 0.5$. MSE as function of the tail parameter, α	70
3.8	Data set for clustering example 1: Cauchy distributed samples with cluster centers $[-6,2]$, $[-2,-2]$, $[2,4]$ and $[3,0]$	75
4.1	Example of a signal corrupted by a single outlier. (a) Linear projections in the noiseless case. (b) Linear projections when the signal is corrupted with a single impulse. (c) Original sparse signal. (d) Reconstructed sparse signal from linear projections using BP with L_2 constraint.	84
4.2	Example of measurements corrupted by a single outlier. (a) Linear projections in the noiseless case. (b) Linear projections corrupted with one impulse. (c) Original sparse signal. (d) Reconstructed sparse signal using BP with L_2 constraint.	86
4.3	Outlier rejection example. (a) Original sparse signal. (b) Reconstructed signal from myriad projections, R-SNR=32.2 dB. (c) Reconstructed signal from linear projections, R-SNR=-28.6 dB.	102
4.4	Comparison results between linear projections and myriad projections for the noiseless case, showing reconstruction SNR as a function of the linearity parameter, K . OMP and BP are used as reconstruction algorithms. The preceding M indicates that the reconstruction is performed using myriad projections.	103
4.5	Reconstruction SNR as a function of the linearity parameter K for impulsive noise models. (a) Additive noise: contaminated p -Gaussian with p varying from 0.001 to 0.1. (b) Additive noise: α -S with α varying from 0.5 to 2.	105

4.6	Myriad measurements performance comparison between optimal K and the proposed estimate for K . Normalized average MSE between myriad projections and clean linear projections for standard Cauchy noise. The scale parameter is varied from 10^{-2} to 10. The normalized MSE of corrupted linear projections is plotted for comparison.	106
4.7	Comparison of linear projections with myriad projections for impulsive observation noise. (a) Contaminated p -Gaussian, R-SNR as a function of the contamination parameter, p . (b) α -S noise, R-SNR as a function of the tail parameter, α . OMP and BPD are used as reconstruction algorithms in both cases. The preceding M indicates that the reconstruction is performed using myriad projections.	108
4.8	Example of a 256×256 image corrupted with salt and pepper noise with density 0.01. (a) Original image. (b) Noisy image. (c) Reconstructed image from linear projections using with BPD, R-SNR=11 dB. (d) Reconstructed image from myriad projections using BPD, R-SNR=23 dB.	110
4.9	Reconstruction SNR as a function of the number of measurements. (a) Linear projections based OMP with Gaussian observation noise. (b) Myriad-based OMP in the noiseless case and α -stable observation noise with α varying from 2 to 0.5.	111
4.10	Outlier rejection example. (a) Original sparse signal (b) Reconstructed signal using Lorentzian BP SNR=115.1 dB (c) Reconstructed signal using OMP SNR=-8.4 dB.	113
4.11	Reconstruction SNR as a function of γ . (a) Effect of the noise strength, standard Cauchy noise with variable scale parameter σ . (b) Effect of the noise impulsiveness, α -stable noise with variable tail parameter α and fixed scale parameter $\sigma = 0.1$	115
4.12	L_2 reconstruction error of Lorentzian BP, before and after debiasing for different Cauchy environments. The theoretical upper bound is plotted for comparison.	116

4.13	Comparison of Lorentzian BP with BPD and OMP in different Cauchy environments. Reconstruction SNR as a function of the scale parameter σ	117
4.14	Comparison of Lorentzian BP with BPD and OMP for impulsive contaminated samples. (a) Contaminated p -Gaussian, $\sigma^2 = 0.01$. R-SNR as a function of the contamination parameter, p . (b) α -S noise, $\sigma = 0.1$. R-SNR as a function of the tail parameter, α	118
4.15	Reconstruction SNR as a function of the number of measurements.	119
5.1	Probability of successful recovery as function of the sparsity level k (noiseless case). $m = 200$	134
5.2	Reconstruction SNR as function of the number of samples m (Gaussian sampling noise, $\sigma^2 = 10^{-2}$). Gaussian distributed non-zero coefficients, $\sigma_x = 10$ and $k = 10$	135
5.3	Reconstruction SNR as function of the number of samples m . ECG signals using CMFB, $M = 16$ and $n = 1024$	136
5.4	Image model example. (a) Original image, (b) Wavelet coefficient histogram with Laplacian distribution fit (dashed) and Meridian distribution fit (blue).	137
5.5	PSNR as function of the number of samples m . Average results on 10 256×256 images.	138
5.6	Image reconstruction example with Lena. Top row: $m = 8000$. LBCS (left), PSNR=18.61 dB and GCBCS (right), PSNR=23.81 dB. Middle row: $m = 20000$. LBCS (left), PSNR=25.56 dB and GCBCS (right), PSNR=26.36 dB. Bottom row $m = 32000$. LBCS (left), PSNR=30.36 dB and GCBCS (right), PSNR=32.10 dB.	139
5.7	Comparison of GCBCS for impulsive contaminated samples. (a) Contaminated p -Gaussian, $\sigma^2 = 0.01$. R-SNR as a function of the contamination parameter, p . (b) α -stable noise, $\sigma = 0.1$. R-SNR as a function of the tail parameter, α	141

5.8	Performance of GCBCS-II as the number of measurements varies for synthetic sparse signals. Reconstruction SNR as a function of the number of measurements.	143
6.1	Weight function for $\gamma = 1$. Large deviations have a weight close to zero whilst small deviations have a weight close to one.	153
6.2	Comparison of LIHT with LS-IHT and WMR for impulsive contaminated samples. (a) Contaminated p-Gaussian, $\sigma^2 = 0.01$. R-SNR as a function of the contamination parameter, p. (b) α -stable noise, $\sigma = 0.1$. R-SNR as a function of the tail parameter, α	161
6.3	Performance of LIHT as the number of measurements varies for synthetic sparse signals. Reconstruction SNR as a function of the number of measurements.	163
6.4	Example of a 256×256 image sampled by a random Hadamard ensemble. Top: clean measurements. Bottom: Cauchy corrupted measurements, $\sigma = 1$	164
6.5	Lena image reconstruction example from measurements corrupted by Cauchy noise. (a) Reconstructed image using LS-IHT, R-SNR=-10.7 dB. (b) Reconstructed image using LS-IHT and noise clipping, R-SNR=6.2 dB. (c) Reconstructed image using LIHT, R-SNR=20.5 dB. (d) Reconstructed image from noiseless measurements using LS-IHT, R-SNR=23.9 dB.	165
6.6	Probability of successful recovery as a function of the number of measurements, for different percentages of partially known support.	166
6.7	Decomposition of an ECG signal using CMFB, $M = 16$ and $n = 1024$	167
6.8	Comparison of LIHT, BP, OMP, CoSaMP, rwns- SL_0 and their partially known support versions for ECG signals.	168
6.9	Wavelet decomposition of the camera man image.	169

6.10 Top left: Original 256×256 image. Top right: Best s -term approximation, $s = 6000$, R-SNR=23.9 dB. Reconstruction from $m = 16000$. Bottom left: LIHT, R-SNR=10.2 dB. Bottom right: LIHT-PKS $k = 2000$, R-SNR=20.4 dB. 171

LIST OF TABLES

3.1	Multi-parameter Estimation Results for GCD Process with length N and $(\theta, \sigma, p) = (0, 1, 2)$	59
3.2	Clustering results for GCD processes and α -stable process	74
5.1	Comparison of reconstruction quality between known δ and estimated δ MBCS. Meridian distributed signals, $n = 1000$, $m = 200$. R-SNR (dB).	133

ABSTRACT

Compressed sensing (CS) is an emerging signal acquisition framework that goes against the traditional Nyquist sampling paradigm. CS demonstrates that a sparse, or compressible, signal can be acquired using a low rate acquisition process. Since noise is always present in practical data acquisition systems, sensing and reconstruction methods are developed assuming a *Gaussian* (light-tailed) model for the corrupting noise. However, when the underlying signal and/or the measurements are corrupted by *impulsive* noise, commonly employed linear sampling operators, coupled with Gaussian-derived reconstruction algorithms, fail to recover a close approximation of the signal. This dissertation develops robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise. To achieve this objective, we make use of robust statistics theory to develop appropriate methods addressing the problem of impulsive noise in CS systems. We develop a generalized Cauchy distribution (GCD) based theoretical approach that allows challenging problems to be formulated in a robust fashion. Robust sampling operators, together with robust reconstruction strategies are developed using the introduced GCD framework.

To solve the problem of impulsive noise embedded in the underlying signal prior the measurement process, we propose a robust nonlinear measurement operator based on the weighed myriad estimator. To recover sparse signals from impulsive noise introduced in the measurement process, a geometric optimization problem based on L_1 minimization employing a Lorentzian norm constraint on the

residual error is introduced. Additionally, robust reconstruction strategies that incorporate prior signal information into the recovery process are developed. First, we formulate the sparse recovery problem in a Bayesian framework using probabilistic priors from the GCD family to model the signal coefficients and measurement noise. An iterative reconstruction algorithm is developed from this Bayesian framework. Second, we develop a Lorentzian norm based iterative hard thresholding algorithm capable of incorporating prior support knowledge into the recovery process. The derived algorithm is a fast method capable of handling large scale problems whilst having robustness against impulsive noise.

Analysis of the proposed methods show that in impulsive environments, when the noise possesses infinite variance, a finite reconstruction error is achieved and furthermore these methods yield successful reconstruction of the desired signal. Experimental results demonstrate that the proposed methods significantly outperform commonly employed compressed sensing sampling and reconstruction techniques in impulsive environments, while providing comparable performance in less demanding, light-tailed environments. Simulation results also show that the proposed algorithms with prior signal information require fewer samples than most existing reconstruction methods to yield approximate reconstruction of sparse signals.

Chapter 1

INTRODUCTION

Conventional approaches to sampling signals follow the famous Shannon-Nyquist theorem: the sampling rate must be at least twice the signal's bandwidth (the so called Nyquist rate) to achieve perfect reconstruction. Nearly all practical acquisition systems are based on this classical result. The traditional mode of data acquisition is to first uniformly sample the signal (at or above the Nyquist rate) to achieve perfect reconstruction. Since for wide-band signals this often results in a large amount of data, a subsequent step is to compress the information about the signal to store it, transmit it or simply process it. The drawback with this traditional framework is that the data has to be acquired first and then compressed, making it very *inefficient*, especially if the amount of acquired data is huge as in many modern applications.

Compressed sensing (CS) is a recently introduced signal acquisition framework that goes against the traditional Nyquist sampling paradigm. CS demonstrates that a sparse, or compressible, signal can be acquired using a low rate acquisition process [18, 37, 41, 79, 88]. The fundamental CS premise is that certain classes of signals, such as natural images, have a concise representation in terms of a sparsity inducing basis where most of the coefficients are zero or small, and only few are significant. Consider the following illustrative example. Let $x(t)$ be a discrete time signal of finite length N and suppose $x(t)$ is the sum of two sinusoids of the form

$$x(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t), \quad t = 0, \dots, N - 1. \quad (1.1)$$

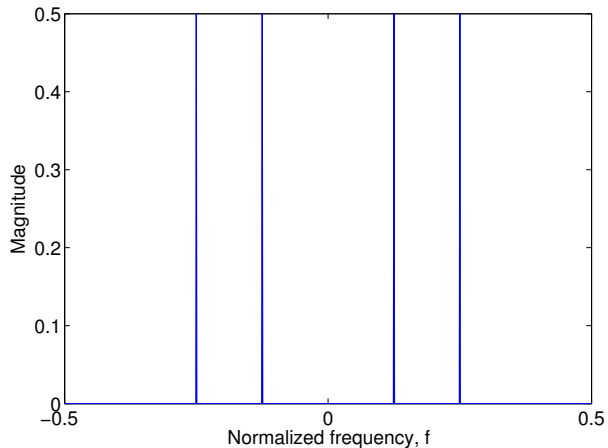


Figure 1.1: Example of sparse signal in the frequency domain. Spectrum magnitude of $x(t)$ for $f_1 = 1/8$ and $f_2 = 1/4$.

The Shannon-Nyquist Theorem dictates that the signal should be sampled at frequency $f_s \geq 2 \max(f_1, f_2)$. Suppose $f_1 = 1/8$ and $f_2 = 1/4$, which indicates that the sampling frequency should be $f_s \geq 1/2$, *i.e.*, we should take at least $N/2$ samples to assure perfect recovery. However, as shown in Fig. 1.1, $x(t)$ is sparse in the frequency domain having only four significant components, which indicates that the Nyquist rate is still high and not close to the intrinsic information rate of the signal. Thus, one question arises from this example: is it possible to sample $x(t)$ in a more efficient fashion, *i.e.* at a lower rate?

In the CS framework, a signal is sampled taking a few linear projections (measurements) onto a set of vectors incoherent with the sparsity basis, and subsequently recovered using an optimization formulation that determines the sparsest representation consistent with the measurements [88]. The key observation is that CS allows us to construct an *efficient* sampling method that captures all the relevant information about the signal, in a signal-independent way, and condense it into a small set of data at once, *i.e.*, sensing and compressing data simultaneously.

From a practical point of view, CS represents a sampling/reconstruction

framework where the signal is sampled close to its intrinsic information rate rather than its Nyquist rate by a simple protocol and then relies on computational power for the recovery process, contrary to current data acquisition-compression methods.

One remarkable aspect of this theory is that it intersects with many areas in pure mathematics, applied and computational mathematics and engineering [37, 40, 47, 79]. As such CS has deep connections with other theories and can be extended to a broad range of applications in practical problems in sciences and engineering. Examples of applications of CS are: data compression, channel coding, inverse problems, image and data acquisition, communications, distributed sensor networks, computer networks, bioinformatics, statistics, approximation theory, medical imaging, astronomy and geosciences. For a review of extensions and applications of CS see for example the resources in [1, 21, 37, 47, 62].

1.1 Motivation

Since noise is always present in real data acquisition systems, a range of different algorithms and methods have been developed that enable approximate reconstruction of sparse signals from noisy compressive measurements [2, 35, 42, 44, 60, 65, 83, 101, 115, 134, 135, 139, 149, 158, 161, 162]. Noise aware algorithms follow three basic approaches: geometric-based algorithms [42, 44, 65, 83, 158], greedy algorithms [134, 135, 161], or complexity based algorithms [2, 60, 115]. Most such algorithms provide bounds for the L_2 reconstruction error based on the assumption that the corrupting noise is bounded, Gaussian, or, at a minimum, has finite variance.

Noise contributions to the overall system can be separated into two models: *observation* noise and *sampling* noise [148]. See Fig. 1.2. Consider first the case of observation noise. Observation noise is any perturbation introduced to the underlying signal *prior* to the sampling process, *e.g.*, channel noise effects in communications or salt and pepper noise in images. The (additive) model of the signal

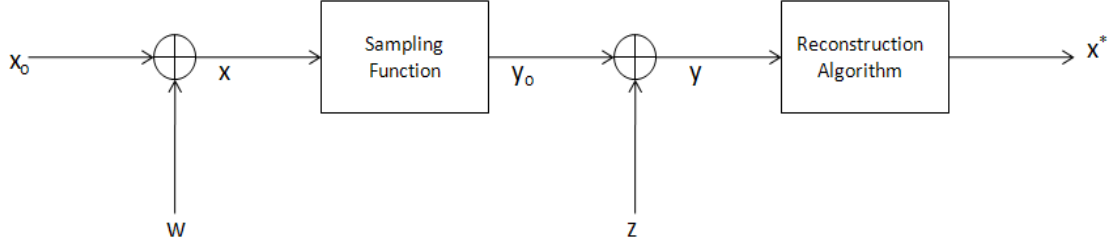


Figure 1.2: Block diagram of noise contributions in compressed sensing.

in this case is:

$$x = x_0 + w, \quad (1.2)$$

where $x_0 \in \mathbb{R}^n$ is the original signal and w is the additive noise. Sampling noise, in contrast, introduces perturbations to the measurements in conjunction with the sampling process, *i.e.*,

$$y = y_0(x) + z \quad (1.3)$$

where $y_0(x) \in \mathbb{R}^m$, $m < n$, is the vector of samples, or measurements of x , and z is the corrupting noise, *e.g.*, quantization noise or sensor noise.

If we consider linear measurements as in the traditional CS literature, then, in the noiseless case, $y = \Phi x_0$, where Φ is the measurement matrix and, for a noisy signal,

$$y = \Phi x = \Phi x_0 + r \quad (1.4)$$

with $r = \Phi w$. When w is Gaussian, r is also Gaussian yielding (1.2) and (1.3) similar; therefore, they can be approached by the same methods and use linear measurements and Gaussian-derived reconstruction algorithms. However, when the signal is corrupted by gross errors or heavy-tailed impulsive noise, linear measurements are

severely degraded, with original signal information masked by large amplitude samples spread throughout the measurements as a result of the linear sampling process. This introduction of large valued corrupting samples, and their spreading across measurements, causes traditional reconstruction algorithms to fail in their attempts to recover a fair approximation of the underlying signal.

Another tenet of traditional reconstruction algorithms that fails in demanding environments is the assumption that the sampling noise has finite variance. If a corrupting process has infinite, or even very large, variance, the allowable, and likely resulting, reconstructions will be far from the desired original signal. Recent works have begun to address the reconstruction of sparse signals from measurements corrupted by outliers, *e.g.*, due to missing data in the measurement process; or in the context of channel coding, due to transmission problems [39,43,125,145]. These works are based on the sparsity of the measurement error pattern to estimate first the error, and then estimate the true signal, in an iterative process. A drawback of this approach is that the reconstruction relies on the error sparsity to first estimate the error, but if the sparsity condition is not met, the performance of the algorithm degrades, or many iterations may be required to yield a fair estimate.

Notably, there exists a broad spectrum of applications where practice has shown non-Gaussian, heavy-tailed processes emerge [121,123]. Examples of such applications are: wireless communications, teletraffic, hydrology, geology, atmospheric noise, economics and image and video processing (see [4,7,24,106,127] and the references therein). Thus, the motivation of this dissertation is the development of robust CS techniques that address these challenging environments.

1.2 Dissertation Objectives and Contributions

The goal of this dissertation is to investigate and develop robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise. In particular we seek the following:

1. To develop a theoretical framework that allows robust formulation of challenging problems in impulsive environments.
2. To develop and analyze robust methods for sampling and reconstructing sparse signals in the presence of impulsive noise using the techniques developed in 1.
3. To develop robust reconstruction strategies that use prior signal information and thus require fewer samples than traditional reconstruction methods to recover sparse signals.

To achieve these objectives, we make use of robust statistics theory to develop appropriate methods addressing the problem of impulsive noise in CS systems. Specifically, we utilize methods derived from the algebraic-tailed Generalized Cauchy distribution (GCD) family [51, 132, 150]. Our contributions are concentrated in three areas:

- *Robust signal processing*: robust estimation and filtering methods, as well as robust error metrics are developed from the GCD family.
- *Compressive sensing methods in impulsive noise*: robust sampling operators, together with robust reconstruction strategies are developed and their properties analyzed.
- *Compressive sensing with prior information*: fast reconstruction strategies that incorporate probabilistic signal models as well as deterministic signal prior information into the recovery process are developed.

The work presented in this dissertation constitutes the foundation of 15 journal and conference articles [48–59, 98, 103, 144], spanning the fields of statistics, estimation, non-linear filtering, sampling theory, sparse approximation and image processing. The contributions in this dissertation can have significant impact in problems where the processes are corrupted by outliers, *e.g.*, missing or saturated

samples. Examples of such problems are: channel coding for erasure channels, real image and data acquisition systems, atmospheric and underwater communications, computer networks, bioinformatics, medical imaging and geosciences.

1.3 Organization

The organization of this document is as follows:

Chapter 2 surveys the theory of compressed sensing. We state the problem of sensing and reconstruction of general signals. Then, we focus on the particular class of sparse signals and formally present the concepts of sparsity and incoherence. We also review the state of the art in reconstruction algorithms for CS and sparse approximation, dividing the discussion in to methods for the ideal noiseless case first, and then robust methods for the realistic noisy scenario.

In Chapter 3, we develop a generalized Cauchy distribution (GCD) based theoretical approach that allows challenging problems to be formulated in a robust fashion. Within this framework, a statistical relationship between the (generalized Gaussian) GGD and GCD families is established. The proposed framework subsumes the GGD based developments (*e.g.*, least squares, least absolute deviation, FLOM, L_p norms, k -means clustering, etc.), thereby guaranteeing performance improvements over traditional problem formulation techniques. The developed theoretical framework includes robust estimation and filtering methods, as well as robust error metrics. A wide array of applications can be addressed through the proposed framework, including, among others: robust regression, robust detection and estimation, clustering in impulsive environments, spectrum sensing when signals are corrupted by heavy-tailed noise and robust compressed sensing methods.

Chapter 4 develops robust methods for sampling and reconstructing sparse signals in the presence of impulsive noise. We approach the problem from a statistical point of view using GCD based robust methods developed in Chapter 3. To solve the problem of impulsive noise embedded in the underlying signal prior the

measurement process, we propose a robust nonlinear measurement operator based on the weighed myriad estimator. To recover sparse signals from impulsive noise introduced in the measurement process, a geometric optimization problem based on L_1 minimization employing a Lorentzian norm constraint on the residual error is introduced. Analysis of the proposed methods show that in impulsive environments, when the noise posses infinite variance, a finite reconstruction error is achieved and furthermore these methods yield successful reconstruction of the desired signal. Simulations demonstrate that the proposed methods significantly outperform commonly employed compressed sensing sampling and reconstruction techniques in impulsive environments, while providing comparable performance in less demanding, light-tailed environments.

Chapter 5 presents reconstruction methods that use prior statistical information in the recovery process. We show that algebraic-tailed impulsive distributions are more suitable models for sparse or compressible signals. Using these models, we formulate the sparse recovery problem in a Bayesian framework using algebraic-tailed priors from the GCD family for the signal coefficients, where the objective is to provide a maximum a posteriori (MAP) signal estimate. This MAP formulation closely resembles L_0 -norm minimization, which features the theoretically lowest bounds on number of measurements required for signal recovery [38]. An iterative reconstruction algorithm is developed from this Bayesian framework. Simulation results show that the proposed method requires fewer samples than most existing reconstruction methods to recover sparse signals, thereby validating the use of GCD priors for the sparse reconstruction problem.

In Chapter 6 we propose a Lorentzian based iterative hard thresholding (IHT) algorithm and a simple modification to incorporate prior signal information in the

recovery process, specifically we study the case of CS with partially known support. The proposed algorithm is a fast method with computational load comparable to the least squares (LS) based IHT, whilst having the advantage of robustness against heavy-tailed impulsive noise. Sufficient conditions for stability are studied and a reconstruction error bound is derived. We also derive sufficient conditions for stable sparse signal recovery with partially known support. Theoretical analysis shows that including prior support information relaxes the conditions for successful reconstruction. Simulations results demonstrate that the Lorentzian-based IHT algorithm significantly outperform commonly employed sparse reconstruction techniques in impulsive environments, while providing comparable performance in less demanding, light-tailed environments. Numerical results also demonstrate that the modifications improve LIHT performance, thereby requiring fewer samples to yield an approximate reconstruction.

Finally, we close in Chapter 7 with concluding remarks and future directions for the work in this dissertation.

Chapter 2

COMPRESSED SENSING BACKGROUND

2.1 Introduction

Nearly all practical signal acquisition systems are based on the classical result of the famous Shannon-Nyquist theorem: the sampling rate must be at least twice the signal's bandwidth (the so called Nyquist rate). The traditional mode of data acquisition is to first uniformly sample the signal (at or above the Nyquist rate) and then process it. However since many modern applications involve wide-band signals that often result in a large amount of data, this process is usually followed by a second step: compression of the information about the signal to reduce the amount of data to be processed. The drawback with this traditional framework is that data has to be acquired first and then compressed, making it very *inefficient*.

This Chapter surveys the theory of compressed sensing (CS), also known as compressive sampling or compressive sensing. CS is a novel framework in mathematics and engineering that goes against the traditional data acquisition paradigm. CS demonstrates that a sparse, or compressible, signal can be acquired using a low rate acquisition process that projects the sparse signal onto a small set of vectors incoherent with the sparsity basis [37, 41, 45, 79]. The fundamental premise is that certain classes of signals, such as natural images, have a concise representation in terms of a sparsity inducing basis where most of the coefficients are zero or small, and only few are significant. The signal is sampled taking a few linear measurements and subsequently recovered using an optimization formulation that determines the

sparsest representation consistent with the measurements. To make this recovery possible, CS relies on two principles: *sparsity* and *incoherence*. Sparsity is related to the rate of information of the signal of interest. Incoherence is related to the sensing modality, specifically to the relationship between sensing functions and the sparsity basis.

From a practical point of view, the importance of CS is that it allows us to construct an *efficient* sampling protocol which captures all the relevant information about the signal and condense it into a small set of data at once; at a rate lower than the Nyquist rate, and in a signal-independent way. In other words, sensing and compressing data simultaneously. The quality of reconstruction depends on the sparsity of the original signal, the reconstruction algorithm and the degree of incoherence. One remarkable aspect of this theory is that is drawn from various subdisciplines in applied mathematics, especially probability theory. In fact the role played by randomness is crucial because one of the most attractive features of CS is that random vectors and randomly selected vectors from orthonormal matrices are incoherent with any sparsity-inducing basis, with high probability [37, 40, 79], therefore allowing easy construction of sensing procedures.

CS interacts with many areas in pure mathematics, applied and computational mathematics and engineering. As such, CS has deep connections with other theories and can be extended to a broad range of applications in practical problems in sciences and engineering. Examples of applications of CS are among others: data compression, channel coding, inverse problems, image and data acquisition, communications, distributed sensor networks, computer networks, bioinformatics, statistics, approximation theory, medical imaging, astronomy, and geosciences. For a review of extensions and applications of CS see, for example, the resources in [1, 21, 37, 47, 62].

The organization of the Chapter is as follows: In section 2.2 we state the problem of sensing and reconstruction of general signals. In section 2.3 we focus on

the particular class of sparse signals and formally present the concepts of sparsity and incoherence. Section 2.4 surveys current reconstruction algorithms for CS and sparse approximation, dividing the discussion in to methods for the ideal noiseless case first, and then robust methods for the realistic noisy scenario. In section 2.5 we establish connections of CS with related fields. Finally, we conclude in section 2.6 with closing thoughts and future directions in the field.

2.2 The Sensing and Reconstruction Problems

In this section we discuss the sampling and reconstruction problems for general classes of signals. In the following we make an abstract treatment and follow the notation in the wavelet community (see for example [71, 116]). Suppose we have an object $x(t)$ (a signal, image or any function of interest) that belongs to a class \mathcal{X} from a Hilbert space \mathcal{H} . We are interested in finding information operators $I_m : \mathcal{X} \rightarrow \mathbb{R}^m$, that sample m pieces of information about x , and reconstruction algorithms $A_m : \mathbb{R}^m \rightarrow \mathcal{X}$ that offer an approximate reconstruction of x from its samples. This approach is rather general and we need more *a priori* information from the class \mathcal{X} to represent it. Suppose there exists an orthonormal basis $\{\psi_n\}_{n \in \mathcal{I}}$ for \mathcal{X} , where $\mathcal{I} \subset \mathbb{Z}$ (extensions to tight frames or redundant dictionaries are immediate). Then any object $x \in \mathcal{X}$ can be represented as

$$x(t) = \sum_{n \in \mathcal{I}} \langle x, \psi_n \rangle \psi_n(t) \quad (2.1)$$

where $\langle \cdot, \cdot \rangle$ represents the inner product in \mathcal{H} . With this representation in hand, *classical sampling theorems* are formulated, where the information operator takes the form

$$I_m(x) = (\langle x, \psi_1 \rangle, \dots, \langle x, \psi_m \rangle) \quad (2.2)$$

and the values $\langle x, \psi_i \rangle$ are referred to as the samples of x . An immediate algorithm A_m to recover x from the samples is to use the series in (2.1). A common example of this sampling/reconstruction strategy is the well known Shannon-Nyquist sampling theorem for the class of band-limited signals.

Theorem 1. *Suppose that $x(t) \in L^1(\mathbb{R})$ and the Fourier transform of x is band-limited to $[-2\pi B, 2\pi B]$. Then,*

$$x(t) = \sum_{k \in \mathbb{Z}} x\left(\frac{k}{2B}\right) \frac{\sin \pi(2Bt - k)}{\pi(2Bt - k)}, \quad (2.3)$$

where the series converges in the L_2 sense.

Observe that the Shannon-Nyquist sampling theorem tells us that a band-limited signal can be sampled uniformly at a rate of at least $2B$ and can be uniquely reconstructed by its samples and the reconstruction formula in (2.3). It is of note that the functions $\sin(\pi t)/\pi t$ are the scale functions of the Shannon wavelets. See [116] for more results in sampling theorems for a broader variety of function spaces. This is a standard set up in information acquisition and in general the underlying signal can be measured through a set of waveforms $\{\phi_i\}_{i=1}^m$ known as sampling kernels. For example if $\{\phi_i\}$ are indicator functions of pixels, the samples are the image data typically collected by a digital camera; or if $\{\phi_i\}$ are complex exponentials or sinusoids we have a collection of Fourier coefficients; for example this modality of sensing is used in magnetic resonance imaging (MRI).

Although the theory can be developed for general infinite dimensional objects, *e.g.*, continuous time/space signals, we restrict our attention to finite dimensional discrete signals $x \in \mathbb{R}^n$ essentially for two reasons: first, it is conceptually simpler; and second, CS discrete theory is far more developed (though some progress in CS for continuous signals has been made, see section 2.6). Having said this, we are interested in *undersampled* situations in which the number m of available samples or

measurements is much smaller than the dimension n of the signal x . Such problems are extremely common in signal processing, communications and in general inverse problems. Consider, for example, a sensor network scenario in which the number of sensors may be limited; or an imaging process via neutron scatter, in which the sensing process is slow and extremely expensive so that the object of interest can only be measured a few times.

These circumstances raise important questions. Is accurate reconstruction possible from $m \ll n$ measurements only? Is it possible to design $m \ll n$ sampling kernels to capture sufficient information about x ? How can we estimate x from these measurements? In principle this is an ill posed problem since we need to solve an underdetermined system of linear equations. Let Φ denote the $m \times n$ sensing matrix with the vectors $\phi_1^*, \dots, \phi_m^*$ as rows, where a^* is the complex transpose of a . The samples can be represented in vector notation as

$$y = \Phi x. \tag{2.4}$$

The problem has infinite solutions \tilde{x} for which $y = \Phi \tilde{x}$ (note the role played by the null space of Φ). However if we explode the a priori information of the class of signals of interest the solution set can be narrowed down to a unique solution. In the case of discrete band-limited signals, the sampling waveforms can be Dirac deltas; and the Shannon–Nyquist theorem tells us that only a few uniformly spaced samples are needed to exactly reconstruct the signal, given that the signal has a very low bandwidth. Here we are interested in a much broader class of signals: sparse or compressible signals.

2.3 Incoherence and Sampling of Sparse Signals

This section presents two fundamental premises in CS: sparsity and incoherence. Sparsity is related to the rate of information or compressibility of the sampled

signal; and incoherence relates to the relationship between the sensing vectors and the sparsity basis. As mentioned in the introduction the quality of reconstruction of CS systems relies on the sparsity of the signal and the incoherence of the sensing vectors and the sparsity basis, therefore the need to dedicate a section for these two concepts.

2.3.1 Sparsity

Many natural signals have concise representation when expressed in a convenient basis. Mathematically speaking, we have a vector $x \in \mathbb{R}^n$ which we expand in an orthonormal basis $\Psi = [\psi_1 \ \psi_2 \ \cdots \ \psi_n]$ as follows

$$x = \sum_{i=1}^n \theta_i \psi_i, \quad (2.5)$$

where θ is the transform coefficient sequence of x , $\theta_i = \langle x, \psi_i \rangle$. For the sake of simplicity we express (2.5) in vector notation as $x = \Psi^T \theta$. The signal is strictly sparse only if s of its coefficients are nonzero, where $s \ll n$. We refer to this type of signals as s -sparse signals. On the other hand, a signal is compressible if $x \in L^p$, *i.e.*

$$\|\theta\|_p = \left(\sum_{i=1}^n |\theta_i|^p \right)^{1/p} \leq R \quad (2.6)$$

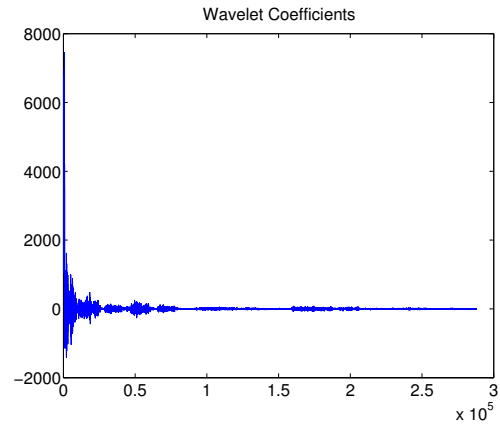
for some $0 < p < 2$ and $R > 0$. In other words, its ordered set of coefficients decay rapidly. Denote by θ_s the vector obtained by only keeping the s largest coefficients of θ , then we have

$$\|\theta - \theta_s\|_2 = \zeta_{2,p} \cdot \|\theta\|_p \cdot (s+1)^{1/2-1/p} \quad (2.7)$$

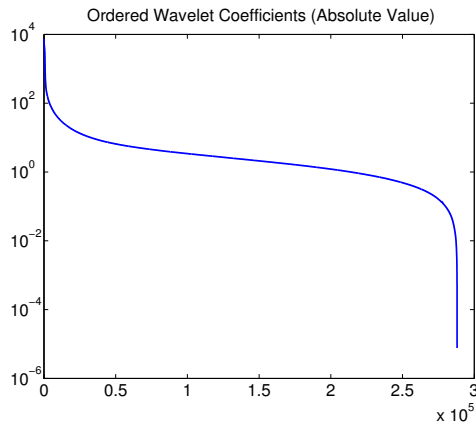
for $s = 1, 2, \dots$, with constant $\zeta_{2,p}$ depending only on p [79]. Thus, for example to approximate θ with error ϵ , we need to keep only the $s \asymp \epsilon^{2p/(p-2)}$ biggest terms in θ . Since Ψ is an orthonormal basis, we have $\|\theta - \theta_s\|_2 = \|x - x_s\|_2$, where $x_s = \Psi^T \theta_s$; and then x is well approximated by x_s . We refer to this class of signals



(a)



(b)



(c)



(d)

Figure 2.1: Image sparsity example. (a) Original 512×512 image with 8 bits per pixel. (b) Wavelet transform coefficients. (c) Ordered wavelet transform coefficients (absolute value) in logarithmic scale. (d) Reconstructed image by zeroing all but the 12.500 largest coefficients.

as s -compressible signals. In plain terms, we can “throw away” a large fraction of the coefficients without much loss [79].

Consider the example shown in Fig. 2.1. Fig. 2.1 (a) shows the original Lena image and its wavelet transform in Fig. 2.1 (b). Although nearly all the image pixels have nonzero values, the wavelet coefficients offer a concise representation: most coefficients are small, if nonzero, and the large coefficients capture most of the image information. Fig. 2.1 (c) presents the ordered wavelet transform coefficients (in absolute value) in logarithmic scale and shows that its ordered set of coefficients decay rapidly. Figure 4.1 (d) shows an example where the perceptual loss is hardly noticeable from the image in Figure 4.1 (a) to its approximation obtained by only keeping 5% of the coefficients.

Many classes of signals and images in nature obey this property and this is the primary reason for the success of standard compression tools based on transform coding [87]. In fact, this principle is what underlies most modern lossy coders such as JPEG-2000 [156] and many others, since a simple method for data compression is to compute θ and then, *adaptively*, encode the locations and values of the s most significant coefficients. Such process requires knowledge of all the n coefficients of θ ; because the locations of the most significant pieces of information may not be known in advance, since they are signal dependent. Sparsity is a fundamental modeling tool which permits efficient signal processing; *e.g.*, accurate statistical estimation and classification, efficient data compression, etc; and therefore an important a priori information to construct *nonadaptive* sampling schemes.

2.3.2 Incoherent sampling

This subsection presents nonadaptive sampling schemes, which are possible thanks to the sparsity property of most natural signals. Let $x \in \mathbb{R}^n$ be a signal that is either sparse or compressible. Suppose we are given the pair (Ω, Ψ) of orthonormal basis of \mathbb{R}^n . The first basis Ω is used to measure x and the second one is used to

represent it. The restriction to orthogonal basis is not essential. The essential premise is that these two bases are incoherent. By incoherent we mean that none of the vectors in Ω have a sparse or compressible representation in the sparsity basis Ψ [40]. Let formally define the mutual coherence.

Definition 1. *The mutual coherence between the sensing basis Ω and the sparsity basis Ψ is*

$$\mu(\Omega, \Psi) = \sqrt{n} \max_{1 \leq j, k \leq n} |\langle \omega_j, \psi_k \rangle|. \quad (2.8)$$

The coherence measures the largest correlation between any two elements of Ω and Ψ , see [84] for further details on the definition of the mutual coherence. If Ω and Ψ contain correlated vectors, the coherence is large, otherwise it is small. Since Ω and Ψ are orthonormal systems, it follows from linear algebra that $\mu(\Omega, \Psi) \in [1, \sqrt{n}]$.

Compressed sensing is mainly concerned with low coherence pairs and in the following we give few examples of such pairs of systems. In our first example, $\Omega = I$ the canonical base of \mathbb{R}^n and Ψ is the Fourier basis. Since Ω is the sensing matrix, this scheme corresponds to the classical sampling procedure in time or space. The time-frequency pair obeys $\mu(\Omega, \Psi) = 1$, *i.e.*, they have *maximal incoherence* [41]. Further, the canonical basis and the Fourier basis are maximally incoherent not only in one dimension but in any dimension. Our second example takes wavelets basis for Ψ and noiselets [68] for Ω . The coherence between Haar wavelets and noiselets is $\sqrt{2}$ and between noiselets and Daubechies D4 and D8 wavelets is about 2.2 and 2.9, respectively, across a wide range of sizes. Noiselets are also maximally incoherent with the canonical and Fourier basis. These results also hold for higher dimensions. The third and final example concerns with random matrices as Ω . Random matrices are largely incoherent with any fixed basis Ψ . If Ω is an orthogonal basis selected uniformly at random then with high probability, the coherence between Ω and Ψ is about $\sqrt{2 \log n}$. By extension matrices with *i.i.d.* entries, *e.g.*, Gaussian or ± 1

Bernoulli entries with normalized columns (in the L_2 sense), exhibit also a low coherence with any fixed Ψ .

Ideally we would like to measure all the n coefficients of x , but we only get to observe a subset of these collected data. Let R be the $m \times n$ matrix that randomly samples m rows of Ω . Then the measurement vector can be written as

$$y = R\Omega x = \Phi x \tag{2.9}$$

where $\Phi = R\Omega$. This approach of random incoherent undersampling was taken in the seminal work of [41] for spectrally sparse signals and showed that the original signal can be recovered with high probability and with a practical recovery algorithm (see next section for reconstruction algorithms). Other works with similar results for this problem, but using different ideas for the proof, are [99, 105, 165]. Later, developments for random matrices were made in parallel by Donoho [79] and Candès [45] where they show conditions that sensing matrices should obey in order to reconstruct the original signal. In [40], the authors extend the concept of random sampling to general orthonormal basis, not only Fourier ensembles or random matrices, where the sensing transforms can be applied quickly and without storing the sensing matrix.

To summarize, in general the sampling procedure is made by taking projections of x on to the set $\{\phi_i\}_{i=1}^m$. The measurement process is a linear map $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, where $m < n$ then $y = \Phi x$ is the vector containing all the measurements. If we set $\Xi = \Phi\Psi^T$ then the measurement vector becomes $y = \Xi\theta$. The vectors $\{\phi_i\}_{i=1}^m$ can be random vectors with independent entries or vectors randomly chosen from an orthogonal basis. The measurements are the only information that we have to reconstruct the original signal and although this is an ill posed problem, it has been shown that we can recover the original signal with an overwhelming probability using the sparsity and incoherence concepts.

2.4 Reconstruction Methods

Compressed sensing algorithms fall within three main categories: geometric approaches, in which geometric constraints are used to find the solution; greedy approaches, which iteratively look for the sparsest solution that best explains the samples; and complexity-based approaches, in which combinatorial optimizations are solved to estimate the original signal. Instead of dividing this section into these three categories we divide the reconstruction algorithms in two cases: the ideal noiseless case, where the system is noise free; and the noisy measurements case, which is a more realistic scenario since noise and perturbations are allowed in the system. Then at each stage we point out to which approach the algorithms belong to.

2.4.1 Notation

Let x be a signal in \mathbb{R}^n and r be a positive integer. We write x_r for the signal in \mathbb{R}^n that is formed by restricting x to its r largest-magnitude components and zero elsewhere. We write $|T|$ to denote the cardinality of the T . If T is a subset of $\{1, 2, \dots, n\}$, then the restriction of the signal to the set T is defined as

$$x_T = \begin{cases} x_i, & i \in T \\ 0, & \text{otherwise.} \end{cases}$$

We denote Φ_T as the the column submatrix of Φ whose columns are listed in the set T . We also write Φ^\dagger to define the pseudoinverse of a tall, full-rank matrix Φ . We keep this notation throughout the rest of this dissertation.

2.4.2 Reconstruction in the Noiseless Case

We start by the ideal noiseless case in which neither the signal nor the measurements are corrupted. In the following we assume, without loss of generality,

that $\Psi = \text{I}$, the canonical basis for \mathbb{R}^n , then $x = \theta$. The ideal algorithm to recover x from the measurements is

$$\min_x \|x\|_0 \text{ subject to } \Phi x = y \quad (2.10)$$

(*i.e.*, find the sparsest vector x such that is consistent with the measurements). The problem in (2.10) is combinatorial and almost intractable; however, it can be relaxed in to a convex problem if the measurement matrices Φ satisfy certain conditions. The convex relaxation is

$$\min_x \|x\|_1 \text{ subject to } \Phi x = y, \quad (2.11)$$

which can be recast as a linear program (LP) and solved efficiently by linear programming techniques. A question arising is: in which cases (2.10) is equivalent to (2.11)? Recent progress in CS resulted in proving the existence of matrices Φ with certain good properties such that the solution of (2.11) is the desired signal. We begin with the results of Candès and Tao in reconstruction from incoherent sampling. It was shown in [45] that if Φ is made of randomly selected rows of an orthogonal system then the recovery condition depends on the mutual coherence μ between the sensing basis and the sparsity basis.

Theorem 2. *Fix $x \in \mathbb{R}^n$ and suppose that x is s -sparse in Ψ . Select m measurements uniformly at random in the sensing domain. Then if*

$$m \geq C\mu^2 s \log n. \quad (2.12)$$

for some positive constant C , the solution to (2.11) is exact with overwhelming probability.

We now turn to the results of Candès *et al.* from a series of works [38, 40, 41,

43, 45]. We first state the definition of the Restricted Isometry Property (RIP) [43] of a sensing matrix Φ .

Definition 2. For every integer $1 \leq s \leq n$ define the s -restricted isometry constant of Φ , δ_s , as the smallest positive quantity such that

$$(1 - \delta_s)\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1 + \delta_s)\|v\|_2^2 \quad (2.13)$$

holds for all $v \in \Omega_s$, where $\Omega_s = \{v \in \mathbb{R}^n \mid \|v\|_0 \leq s\}$.

A matrix Φ is said to satisfy the RIP of order s if $\delta_s \in (0, 1)$. This property requires that every set of columns with cardinality less than s , approximately behaves like an orthonormal system. When this property holds, Φ approximately preserves the Euclidean length of s -sparse vectors, which it implies that s -sparse vectors can not be in the null space of Φ . It was shown in [43] that if Φ has restricted isometry constants such that $\delta_s + \delta_{2s} + \delta_{3s} < 1$, then solving (2.11) recovers any sparse signal with support size of at least s . It is also shown that random matrices with Gaussian or sub-Gaussian entries have restricted isometry constants in the interval $[0, 1)$ with high probability provided that $m = O(s \log(n))$ [19].

The main idea behind the RIP is that the sensing matrix Φ behaves like an orthonormal system for sparse signals and approximately preserves the Euclidean norm. Stating this in other words, the matrix Φ is designed to preserve as much information as possible about x in the projections y despite of the dimensionality reduction. The RIP has gained wide acceptance in the signal processing community due to its simplicity for proving arguments in sparse reconstruction. There has been recent work on constructing CS matrices that obey the RIP, see for example [19, 130] for random matrices results and [77] for deterministic constructions.

The above results were refined in [38] for recovery of more general s -compressible signals.

Theorem 3. *Assume that $\delta_{2s} < \sqrt{2} - 1$. Then the solution x^* to (2.11) obeys*

$$\begin{aligned} \|x^* - x\|_2 &\leq C_0 \cdot \|x - x_s\|_1 / \sqrt{s} \quad \text{and} \\ \|x^* - x\|_1 &\leq C_0 \cdot \|x - x_s\|_1 \end{aligned} \tag{2.14}$$

for some constant C_0 , where x_s is the vector x with the largest s coefficients kept and the rest set to 0.

The conclusions of theorem 3 are stronger than the previous results, since now if x is strictly s -sparse, then $x^* = x$. But if x is not s -sparse, but is compressible, then the quality of reconstruction is almost the same as if we would have known in advance the locations of the s largest coefficients. See also [67] for results in the same direction.

The use of L_1 norm as a sparsity-encouraging function traces back several decades. A first application was used in reflection-seismology, in which a sparse reflection, indicating meaningful changes between subsurface layers, was sought from limited data [66, 153]. The algorithm is now known as Basis Pursuit (BP) and was previously used to find sparse representations on over complete dictionaries [65, 82] and recently used to find sparse solutions for underdetermined systems of linear equations [80]. All the aforementioned algorithms are geometric approaches since they use geometric constraints and geometric optimization tools to find the solution.

Other approach used to find a sparse solution is the use of greedy algorithms that iteratively construct a sparse approximation for the signal. These type of algorithms include Matching Pursuit (MP) [128] and Orthogonal Matching Pursuit (OMP) [161]. Matching Pursuit is a greedy algorithm that iteratively incorporates in the reconstructed signal the component from the measurement set that explains the largest portion of the residual from the previous iteration. The stop criterium is reached when the residual reaches a magnitude below a set threshold. The algorithm

is described next. Denote ϕ_k as the k -th column vector of Φ and $\delta_j \in \mathbb{R}^n$ as the Kronecker delta at position j . Set the residual at time zero as $r^{(0)} = y$ and the initial solution as $x^{(0)} = 0$ (the n -dimensional zero vector). The algorithm is described in Algorithm 1.

Algorithm 1 MP Algorithm

Require: Sensing matrix Φ and measurements y .

- 1: Initialize $i = 0$, $\hat{x}^{(0)} = 0$ and $r^{(0)} = y$.
 - 2: **while** halting criterion **do**
 - 3: $i \leftarrow i + 1$
 - 4: $c_k^{(i)} = \langle \phi_k, r^{(i-1)} \rangle$
 - 5: $\hat{k} = \arg \max_k |c_k^{(i)}|$
 - 6: $\hat{x}^{(i)} = \hat{x}^{(i-1)} + c_{\hat{k}}^{(i)} \delta_{\hat{k}}$
 - 7: $r^{(i)} = r^{(i-1)} - c_{\hat{k}}^{(i)} \phi_{\hat{k}}$
 - 8: **end while**
 - 9: **return** \hat{x}
-

The procedure is repeated until $\|r^{(i)}\|_2 \leq \epsilon$, for some predetermined $\epsilon > 0$. MP was developed in the statistics community under the name of Projection Pursuit Regression [102]. It was introduced to the signal processing community by [128] and independently by [146]. In the approximation community, MP is known as the Pure Greedy Algorithm [157]. For a deeper review of this algorithm see [157].

2.4.3 Reconstruction from Noisy Measurements

Compressed sensing systems are not immune to noise contributions due to sensor noise, finite precision or quantization effects; therefore in order to be really powerful, CS needs to be able to deal with such perturbations; *i.e.*, at the very least, small perturbations in the data should cause small perturbations in the reconstruction. Noise contributions to the overall system can be separated into two models: *observation* noise and *sampling* noise [148]. Consider first the case of observation noise. Observation noise is any perturbation introduced to the underlying signal

prior the sampling process, *e.g.*, noisy channel effects or salt and pepper noise in images. The (additive) model of the signal in this case is:

$$x = x_0 + w, \tag{2.15}$$

where $x_0 \in \mathbb{R}^n$ is the original signal and w is the additive noise. Sampling noise, in contrast, introduces perturbations to the measurements in conjunction with the sampling process, *i.e.*,

$$y = y_0 + z \tag{2.16}$$

where $y_0 \in \mathbb{R}^m$, $m < n$, is the vector of samples, or measurements of x , and z is the corrupting noise, *e.g.*, quantization noise or sensor noise.

If we consider Φ as the linear measurement operator, then the overall noise contribution is $e = z + \Phi w$. When w and z are both Gaussian, e is also Gaussian and both problems can be simultaneously addressed by the same reconstruction techniques [115]. In the presence of noise, variations of the aforementioned algorithms have been shown to reliably approximate the signal, assuming that certain apriori information is known about the signal and/or the noise process [42, 83, 115, 161]. These reconstruction methods are intimately related to estimation theory, since they all try to make an estimate of x_0 from the (possibly corrupted) measurements y . Most of the algorithms used in compressed sensing solve one of the following formulations.

We begin with geometric approaches. The basic approach is Basis Pursuit with L_2 constraint, which relaxes the requirement that the reconstructed signal explain exactly the measurements [42, 83, 163]. Instead, the constraint is expressed in terms of the maximum noise level that we can tolerate, *i.e.*, the maximum distance of the noisy measurements from the re-measured reconstructed signal. The

reconstruction solves the optimization problem

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|y - \Phi x\|_2 \leq \epsilon, \quad (2.17)$$

for some small $\epsilon > 0$. This is a convex problem, in fact a second-order cone program (SOCP) and can be solved efficiently. In [42] it is shown that if x is s -sparse, the noise is power-limited to ϵ and if $\delta_{3s} + 3\delta_{4s} < 2$; then the reconstructed signal, \hat{x} , is guaranteed to be within $C\epsilon$ of the original signal x , $\|x - \hat{x}\|_2 \leq C\epsilon$, where the constant C depends on δ_{3s} and δ_{4s} (measurement parameters) and not on the noise level. In [38] this result is refined for general compressible signals.

Theorem 4. *Assume $\delta_{2s} < \sqrt{2} - 1$. Then the solution to (2.17) obeys*

$$\|x - \hat{x}\|_2 \leq C_0 \cdot \|x - x_s\|_1 / \sqrt{s} + C_1 \cdot \epsilon \quad (2.18)$$

for some positive constants C_0 and C_1 that depend on δ_{2s} .

The first term in the right hand side of the bound is the reconstruction error that we have for the noiseless case and the second term is just proportional to the noise level. The constants C_0 and C_1 are typically small. Unfortunately, in practice, noise is not necessarily power limited and, even when it is, the power limit is usually unknown. The problem in (2.17) was proposed in [153] and is widely used in sparse estimation problems with alternative formulations.

The Lasso and basis pursuit de-noising (BPD) are two alternative formulations of the same objective in (2.17). Basis pursuit de-noising was used in the context of wavelet de-noising and statistical regression [65]. It converts the SOCP in (2.17) in to the unconstrained convex problem:

$$\hat{x} = \arg \min_x \|x\|_1 + \lambda \|y - \Phi x\|_2^2. \quad (2.19)$$

With appropriate parameter correspondence, this formulation is equivalent to the Lasso [158]:

$$\hat{x} = \arg \min_x \|y - \Phi x\|_2^2 \text{ subject to } \|x\|_1 \leq q. \quad (2.20)$$

Furthermore, is demonstrated in [158] that as λ ranges from zero to infinity, the solution path of (2.19) is the same as the solution path of (2.20) as q ranges from infinity to zero. It follows that determining the proper value of λ is akin to determining the power limit of the noise. Least angle regression (LAR) is a more general model selection algorithm that contains the solution to the Lasso as a particular case and is shown to converge in a few steps [93]. Recently, fast algorithms have been developed to find the solution of BPD for large scale systems [101, 122].

The Dantzig Selector is an alternative convex program for model selection, useful when $m \ll n$ and the noise is Gaussian with bounded variance σ [44]. Let $\delta_{2s} + \delta_{3s} < 1$ and m be sufficiently large, then the program

$$\min_x \|x\|_1 \text{ subject to } \|\Phi^T(y - \Phi x)\|_\infty \leq \sigma \sqrt{2 \log n} \quad (2.21)$$

reconstructs a signal that satisfies

$$\|x - \hat{x}\|_2 \leq C\sigma \sqrt{2s \log n}. \quad (2.22)$$

The results not only apply for strict sparse signals but also to compressible signals $x \in L^p$ for $p < 2$. This algorithm requires the a priori knowledge of the error variance and furthermore a finite variance for a bounded reconstruction error.

All of the above reconstruction formulations are based on the same principle: minimizing the L_1 norm, under certain conditions on the measurement matrices and sparsity basis, will find the solution with the sparsest representation and furthermore will find the location of the non-zero coefficients of the sparse representation.

They all rely on having noise with finite and small variance to perform a fair reconstruction. These algorithms are often followed by a subsequent step, known as de-biasing, in which a standard least squares problem is solved on the support, i.e. find x that solves

$$\min \|y - \Phi_I x\|_2^2 \quad (2.23)$$

where $I = \{i : |\hat{x}_i| > \alpha\}$ for some threshold α , \hat{x} is the solution found for the L_1 algorithm. The estimated signal after the de-biasing process, \tilde{x} , is defined as $\tilde{x}_i = x_i \forall i \in I$ and $x_i = 0 \forall i \in I^c$.

Following the same geometric approach, recent works have shown that non-convex reconstruction formulations can recover a sparse signal with fewer measurements than current geometric methods, while preserving the same reconstruction quality [46, 63, 73, 152]. These approaches are based on approximating the L_0 norm (quasi-norm) with a sequence of continuous functions $\{f_\sigma\}$, that converge to the L_0 norm as $\sigma \rightarrow 0$ or $\sigma \rightarrow \infty$ in some sense. In [63], the authors replace the L_1 norm in BP with the L_p norms, with $0 < p < 1$ to approximate the L_0 norm and encourage sparsity in the solution. They show RIP for matrices that preserve the L_p norm instead of only the Euclidean norm (L_2). The work in [152] extend the ideas of L_p norms to the noisy case. In [73], Daubechies *et al.* show how an iteratively re-weighted least squares approach, based on the FOCUSS algorithm [110], can find an sparse solution. The idea is that giving a large weight to small components encourages sparse solutions. Following the same philosophy, Candés *et al.* use a re-weighted L_1 minimization approach to find a sparse solution [46].

We now turn to greedy approaches and focus on Orthogonal Matching Pursuit (OMP) since other greedy algorithms share the same philosophy. OMP is based on MP but additionally orthogonalizes the residual against all measurement vectors selected in previous iterations [161]. By doing so the performance of the algorithm is improved and provides better reconstruction compared to plain MP although the

complexity is increased. The algorithm is described in Algorithm 2.

Algorithm 2 OMP Algorithm

Require: Sensing matrix Φ and measurements y .

- 1: Initialize $i = 0$, $\hat{x}^{(0)} = 0$, $\Lambda = \emptyset$ and $r^{(0)} = y$.
 - 2: **while** halting criterion **do**
 - 3: $i \leftarrow i + 1$
 - 4: $e = \Phi^T r^{(i)}$
 - 5: $\Omega = \arg \max_j |e(j)|$
 - 6: $\Lambda = \Omega \cup \text{supp}(\hat{x}^{(i-1)})$
 - 7: $\hat{x}^{(i+1)} = \Phi_{\Lambda}^{\dagger} y$
 - 8: $r^{(i)} = y - \Phi_T \hat{x}^{(i)}$
 - 9: **end while**
 - 10: **return** \hat{x}
-

The procedure is repeated until $\|r^{(i)}\|_2 \leq \epsilon$, for some predetermined $\epsilon > 0$, or if the sparsity level of the signal, s , is known, then perform only s iterations of the algorithm. The number of measurements required for OMP is also $O(s \log n)$ if random sensing matrices are used [161]. The orthogonalization step is similar to the debiasing step defined in (2.23) and is performed at every iteration of the algorithm except at the end. Note that OMP never selects the same atom (column of Φ) twice because the residual is orthogonal to the atoms that have already been chosen. Both MP and OMP have been shown to converge to a solution that fully explains the data and the noise; however, only OMP is guaranteed to converge to a sparse solution [161]. Experiments have shown that proper termination of the algorithm is a practical way to reject the measurement noise in the reconstruction. However, the conditions for proper termination involve the knowledge of the sparsity of the signal or requires the knowledge of the noise level ϵ to apply the stopping condition $\|r^{(i)}\|_2 \leq \epsilon$. Furthermore, it requires the noise to have bounded and small variance to achieve a good performance.

OMP was developed independently by many researchers. The earliest reference appears to be a 1989 paper of Chen, Billings and Luo [64]. The first signal

processing works on OMP arrived around 1993 [76, 143]. OMP was later proposed as an algorithm for sparse signal approximation over redundant complete dictionaries in [160] and its connections with CS were made in [161] to recover sparse signals from random measurements. In [135], the authors make a bridge between geometric algorithms and greedy algorithms, providing an iterative algorithm with the ease of implementation of OMP and the theoretical guarantees of the geometric approaches. The algorithm is known as regularized OMP. Another greedy approach similar in spirit to OMP is CoSaMP [134], which also offer strong theoretical guarantees and performs faster than most of the current algorithms. See [134] for a complete comparison of current greedy algorithms and their limitations.

Another greedy approach used for the reconstruction problem is the iterative hard thresholding algorithm (IHT). The IHT algorithm is a simple iterative method that does not require matrix inversion at any point and provides near-optimal error guarantees [33, 34]. The algorithm is described as follows.

Let $x^{(t)}$ denote the solution at iteration time t and set $x^{(0)}$ to the zero vector. At each iteration t the algorithm computes

$$x^{(t+1)} = H_s(x^{(t)} + \mu\Phi^T(y - \Phi x^{(t)})), \quad (2.24)$$

where $H_s(a)$ is the non-linear operator that sets all but the largest (in magnitude) s elements of a to zero and μ is a step size. If there is no unique set, a set can be selected either randomly or based on a predefined ordering. Convergence of this algorithm is proven in [32] and a theoretical analysis for compressed sensing problems is presented in [33, 34].

The third approach is a complexity-based approach that solves iteratively a combinatorial optimization problem [115]. In the following we make a brief description of the idea behind this approach (since it is not really an algorithm). Suppose Φ is made of random entries of variance $\mathbb{E}(\phi_{ij}^2) = 1/n$ and that the noise e is formed by

i.i.d Gaussian r.v.'s with variance σ^2 , independent of $\{\phi_{ij}\}$. The goal is to construct an estimate of x_0 from the observations y . Suppose that x_0 is a compressible signal such that

$$\frac{\|x_0 - x_s\|_2^2}{n} \leq C_A s^{-2\alpha} \quad (2.25)$$

for some $C_A > 0$ and $\alpha \geq 0$. Suppose we have a countable collection \mathcal{X} of candidate reconstruction signals and suppose also that $\|x\|_2^2 \leq nB^2$, for every $x \in \mathcal{X}$ and for some $B > 0$. Select a reconstruction signal according to

$$\hat{x}_m = \arg \min_{x \in \mathcal{X}} \left\{ \frac{\|y - \Phi x\|_2^2}{m} + \frac{2 \log 2 \log n \|x\|_0}{m\epsilon} \right\} \quad (2.26)$$

where $\epsilon = 1/(21(B + \sigma)^2)$. Then, Haupt and Nowak prove that there exist positive constants $C_1 = C_1(B, \sigma)$ and $C_2 = C_2(B, \sigma, C_A)$ such that

$$\mathbb{E} \left[\frac{\|x_0 - \hat{x}_m\|_2^2}{n} \right] \leq C_1 C_2 \left(\frac{m}{\log n} \right)^{-2\alpha/(2\alpha+1)}. \quad (2.27)$$

The authors give concrete values for the constants C_1 and C_2 and practical implementation algorithms based on EM. They also give bounds for the reconstruction error similar to those in [38, 42] for Gaussian noise and Rademacher sensing matrices. An important contribution to this approach is the work of [27], where the authors present a unified approach between the geometric and combinatorial approaches, using deterministic sensing matrices and generalizing the notion of RIP from Euclidean norm to general L_p norms.

2.5 Connections of CS with Other Fields

In the following we briefly explore connections of CS with two important areas in applied mathematics: error correction and high-dimensional geometry. The reason to establish these relations is that they share similarities in their problem formulation and therefore ideas from these fields can be applied to CS or viceversa.

The basic problem in these three areas is to reduce the dimension of some high-dimensional vector and try to preserve as much information as possible about it.

We begin with coding theory or error correction theory. Let \mathbb{F} be any arbitrary scalar field. Suppose we want to reliably transmit a vector $x \in \mathbb{F}^M$ through a channel. A frequent approach is to encode the information of x into a vector y of higher dimension, say N . This encoding process can be modeled as $y = Cx$, where C is the $N \times M$ coding matrix or generator matrix. In the decoding process, we have available a matrix B , such that $BC = 0$; B is called a parity check matrix and is any $(N - M) \times M$ matrix whose null space is in the range of C in \mathbb{F}^N . The transmitted information is of the form $y = Cx + e$, where e is the error pattern or error vector (the positions of the errors are unknown but sparse). Applying B to the received vector gives

$$\tilde{y} = B(Cx + e) = Be \tag{2.28}$$

since $BC = 0$. Therefore the decoding problem is reduced to that of recovering the error vector e from the observations Be . This is again an ill posed problem since we have fewer equations than unknowns, but with the assumption that only a fraction of e is contaminated (sparsity), so the relation between CS and coding theory is established. However the reconstruction methods used in one field may not work properly for the other (at least in straight manner). CS generally deals with real fields or complex fields, meanwhile error correction usually deals with finite fields. Having said this; if the vector x belongs to \mathbb{R}^M , CS techniques can be employed to recover x as proposed in [43]. The authors proposed to recover solving

$$\min_x \|y - Cx\|_1 \tag{2.29}$$

which is equivalent to solve

$$\min_d \|d\|_1 \text{ subject to } Bd = Be. \quad (2.30)$$

They prove that if C has *i.i.d.* Gaussian entries, then the decoding is exact, provided the number of errors is less than a certain number that depends on N and M . Other examples of CS in error coding are [70, 168].

Lets now turn to high-dimensional geometry, in which CS has foundations. Donoho and Tanner have results from polytope geometry to obtain very precise estimates about the minimal number of Gaussian measurements needed to reconstruct an s -sparse signals [81, 85, 86]. Let A be a $d \times n$ matrix, $d < n$ and let C be the regular cross polytope (orthoplex) in \mathbb{R}^n . Define P as the projected polytope of C on to the subspace spanned by A . Then, they showed that the minimal number of random measurements is related to the number of faces of P , provided A is a Gaussian matrix.

Another relationship between CS and high-dimensional geometry comes from dimensionality reduction. Traditional techniques for dimensionality reduction are for example PCA, ICA, MDS and their variations but with the drawback that they only capture the signal information for limited cases. In the CS framework the sensing matrix is basically a linear map that projects the original object on to a subspace of lower dimensionality. Since the RIP requires that that the sensing matrix preserves the Euclidean norm up to certain bounds, CS can be used as a powerful tool for dimensionality reduction, detection and estimation of sparse signals. Furthermore the works in [20, 166] extend the use of the CS framework to the broader class of signals of manifold-based models, which arise in both parametric and non-parametric signal families. It is shown there, that random projections are an effective way of performing the difficult task of manifold learning using a lower dimensional space that saves computational resources.

2.6 Concluding Remarks

This Chapter presents a survey of the basic theory behind the now mature field of compressed sensing. We review the powerful tools of CS for signal sampling and signal reconstruction/estimation methods. One note to make is that the early papers on CS, [41, 45, 79], initiated a large and fascinating body of literature in which other ideas and approaches have been proposed (see for example the resources in [1, 21, 37, 47, 62]). Among these new directions, CS with prior information inclusion in the recovery process and CS-based analog-to-information protocols; are the two areas with more promising future in the field.

Recent CS literature has investigated the concept of exploiting *prior* signal information. It is shown that modifying the CS framework to include prior signal knowledge improves the reconstruction results using fewer measurements [22, 90, 104, 119, 164]. For instance, Vaswani *et. al* assume that part of the signal support is known *a priori*, reducing the problem to finding the unknown portion of support and thereby requiring fewer samples to yield an accurate reconstruction [164]. Baraniuk *et. al* introduced a model-based CS theory that reduces the degrees of freedom of a sparse/compressible signal by permitting only certain configurations of large and zero/small signal coefficients [22, 90, 91, 124]. Similarly, a recovery framework based on a structured union of subspaces is proposed by Eldar and Mishali [97], while source statistics, modeled as stochastic processes, are exploited in [15, 61, 92, 104, 120].

The implications of the aforementioned works is that sampling techniques can be implemented with a lower acquisition rate. If better reconstruction algorithms are available, then thinking in data acquisition implementations is not a crazy idea. There have been a tremendous effort in this direction to construct such devices. The work in [89] reports the implementation of a single pixel camera architecture with promising results. From the same group, results of analog-to-information devices were reported in [126, 147].

Finally we would like to close this Chapter stating one of the most challenging problem in CS, which still is an open problem: compressed sensing of continuous signals. As we mentioned earlier CS theory is well developed for discrete signals, although there still are many questions to be answered, but there is no straight extension of this theory to more general infinite dimensional continuous signals. Recently in [96] and [151] the authors present (separately) theoretical results that pave the road to extend the theory of CS to infinite dimensional function spaces. These extensions to general abstract functions is the first step for applying the CS framework to continuous time-space signals but there is still much work to be done and many questions unanswered.

Chapter 3

A GENERALIZED CAUCHY DISTRIBUTION FRAMEWORK FOR PROBLEMS REQUIRING ROBUST BEHAVIOR

3.1 Introduction

Traditional signal processing and communications methods are dominated by three simplifying assumptions: (1) the systems under consideration are linear; the signal and noise processes are (2) stationary and (3) Gaussian distributed. Although these assumptions are valid in some applications and have significantly reduced the complexity of techniques developed, over the last three decades practitioners in various branches of statistics, signal processing, and communications have become increasingly aware of the limitations these assumptions pose in addressing many real-world applications. In particular, it has been observed that the Gaussian distribution is too light-tailed to model signals and noise that exhibits impulsive and non-symmetric characteristics [123]. A broad spectrum of applications exists in which such processes emerge, including wireless communications, teletraffic, hydrology, geology, atmospheric noise compensation, economics, and image and video processing (see [4, 24] and references therein). The need to describe impulsive data, coupled with computational advances that enable processing of models more complicated than the Gaussian distribution, has thus led to the recent dynamic interest in heavy-tailed models.

Robust statistics – the stability theory of statistical procedures – systematically investigates deviation from modeling assumption affects [118]. Maximum likelihood (ML) type estimators (or more generally, M-estimators), developed in the theory of robust statistics are of great importance in *robust signal processing techniques* [121]. M-estimators can be described by a cost function defined optimization problem or by its first derivative, the latter yielding an implicit equation (or set of equations) that is proportional to the influence function. In the location estimation case, properties of the influence function describe the estimator robustness [118]. Notably, ML location estimation forms a special case of M-estimation, with the observations taken to be independent and identically distributed and the cost function set proportional to the logarithm of the common density function.

To address as wide an array of problems as possible, modeling and processing theories tend to be based on density families that exhibit a broad range of characteristics. Signal processing methods derived from the generalized Gaussian distribution (GGD), for instance, are popular in the literature and include works addressing heavy-tailed process [3, 4, 6, 24, 170]. The GGD is a family of closed form densities, with varying tail parameter, that effectively characterizes many signal environments. Moreover, the closed form nature of the GGD yields a rich set of distribution optimal error norms (L_1 , L_2 , and L_p), and estimation and filtering theories, *e.g.*, linear filtering, weighted median filtering, fractional low order moment (FLOM) operators, etc. [4, 6, 8, 25, 154]. However, a limitation of the GGD model is the tail decay rate — GGD distribution tails decay exponentially rather than algebraically. Such light tails do not accurately model the prevalence of outliers and impulsive samples common in many of today’s most challenging statistical signal processing and communications problems [4, 9, 106].

As an alternative to the GGD, the α -stable density family has gained recent popularity in addressing heavy-tailed problems. Indeed, symmetric α -stable

processes exhibit algebraic tails and, in some cases, can be justified from first principles (Generalized Central Limit Theorem) [36, 137, 172]. The index of stability parameter, $\alpha \in (0, 2]$, provides flexibility in impulsiveness modeling, with distributions ranging from light-tailed Gaussian ($\alpha = 2$) to extremely impulsive ($\alpha \rightarrow 0$). With the exception of the limiting Gaussian case, α -stable distributions are heavy-tailed with infinite variance and algebraic tails. Unfortunately, the Cauchy distribution ($\alpha = 1$) is the only algebraic-tailed α -stable distribution that possesses a closed form expression, limiting the flexibility and performance of methods derived from this family of distributions. That is, the single distribution Cauchy methods (Lorentzian norm, weighted myriad) are the most commonly employed α -stable family operators [10, 106–108].

The Cauchy distribution, while intersecting the α -stable family at a single point, is generalized by the introduction of a varying tail parameter, thereby forming the Generalized Cauchy density (GCD) family. The GCD has a closed form pdf across the whole family, as well as algebraic tails that make it suitable for modeling real-life impulsive processes [132, 150]. Thus the GCD combines the advantages of the GGD and α -stable distributions in that it possesses (1) heavy, algebraic tails (like α -stable distributions) and (2) closed form expressions (like the GGD) across a flexible family of densities defined by a tail parameter, $p \in (0, 2]$. Previous GCD family development focused on the particular $p = 2$ (Cauchy distribution) and $p = 1$ (meridian distribution) cases, which lead to the myriad¹ and meridian [7, 9] estimators, respectively. These estimators provide a robust framework for heavy-tail signal processing problems.

In yet another approach, the generalized- t model is shown to provide excellent fits to different types of atmospheric noise [131]. Indeed, Hall introduced the family

¹ It should be noted that the original authors derived the myriad filter starting from α -stable distributions, noting that there are only two closed-form expressions for α -stable distributions [106–108].

of generalized- t distributions in 1966 as an empirical model for atmospheric radio noise [112]. The distribution possesses algebraic tails and a closed form pdf. Like the α -stable family, the generalized- t model contains the Gaussian and the Cauchy distributions as special cases, depending on the degrees of freedom parameter. It is shown in [108] that the myriad estimator is also optimal for the generalized- t family of distributions. Thus we focus on the GCD family of operators, as their performance also subsumes that of generalized- t approaches.

In this Chapter, we develop a GCD based theoretical approach that allows challenging problems to be formulated in a robust fashion. Within this framework, we establish a statistical relationship between the GGD and GCD families. The proposed framework subsumes GGD based developments (*e.g.*, least squares, least absolute deviation, FLOM, L_p norms, k -means clustering, etc.), thereby guaranteeing performance improvements over traditional problem formulation techniques. The developed theoretical framework includes robust estimation and filtering methods, as well as robust error metrics. A wide array of applications can be addressed through the proposed framework, including, among others: robust regression, robust detection and estimation, clustering in impulsive environments, spectrum sensing when signals are corrupted by heavy-tailed noise, and robust compressed sensing (CS) and reconstruction methods. As illustrative and evaluation examples, we formulate four particular applications under this framework: (1) filtering for power line communications, (2) estimation in sensor networks with noisy channels, and (3) fuzzy clustering.

The organization of the Chapter is as follows: In Section 3.2, we present a brief review of M-estimation theory and the generalized Gaussian and generalized Cauchy density families. A statistical relationship between the GGD and GCD is established and the ML location estimate from GCD statistics is derived. An M-type estimator, coined *M-GC* estimator, is derived in Section 3.3 from the cost function

emerging in GCD-based ML estimation. Properties of the proposed estimator are analyzed and a weighted filter structure is developed. Numerical algorithms for multi-parameter estimation are also presented. A family of robust metrics derived from the GCD are detailed in Section 3.4 and their properties are analyzed. Three illustrative applications of the proposed framework are presented in Section 3.5. Finally, we conclude in Section 3.6 with closing thoughts and future directions.

3.2 Distributions, Optimal Filtering and M-Estimation

This section presents M-estimates, a generalization of maximum likelihood (ML) estimates, and discusses optimal filtering from a ML perspective. Specifically, it discusses statistical models of observed samples obeying generalized Gaussian statistics and relates the filtering problem to maximum likelihood estimation. Then, we present the generalized Cauchy distribution and a relation between GGD and GCD random variables is introduced. The ML estimators for GCD statistics is also derived.

3.2.1 M-Estimation

In the M-estimation theory the objective is to estimate a deterministic but unknown parameter $\theta \in \mathbb{R}$ (or set of parameters) of a real-valued signal $s(i; \theta)$ corrupted by additive noise. Suppose we have N observations yielding the following parametric signal model

$$x(i) = s(i; \theta) + n(i) \quad (3.1)$$

for $i = 1, 2, \dots, N$, where $\{x(i)\}_{i=1}^N$ and $\{n(i)\}_{i=1}^N$ denote the observations and noise components, respectively. Let $\hat{\theta}$ be an estimate of θ , then any estimate that solves the minimization problem of the form

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \rho(x(i); \theta) \quad (3.2)$$

or by an implicit equation

$$\sum_{i=1}^N \psi(x(i); \hat{\theta}) = 0, \quad (3.3)$$

is called an M-estimate (or maximum likelihood type estimate). Here $\rho(x; \theta)$ is an arbitrary cost function to be designed and $\psi(x; \theta) = (\partial/\partial\theta)\rho(x; \theta)$. Note that ML-estimators are a special case of M-estimators with $\rho(x; \theta) = -\log f(x; \theta)$, where $f(\cdot)$ is the probability density function of the observations. In general, M-estimators do not necessarily relate to probability density functions.

In the following we focus on the location estimation problem. This is well-founded, as location estimators have been successfully employed as moving window type filters [4,121,154]. In this case, the signal model in (3.1) becomes $x(i) = \theta + n(i)$ and the minimization problem in (3.2) becomes

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \rho(x(i) - \theta) \quad (3.4)$$

or

$$\sum_{i=1}^N \psi(x(i) - \hat{\theta}) = 0. \quad (3.5)$$

For M-estimates it can be shown that the influence function is proportional to $\psi(x)$ [113, 118], meaning that we can derive the robustness properties of an M-estimator, namely efficiency and bias in the presence of outliers, if ψ is known.

3.2.2 Generalized Gaussian Distribution

The statistical behavior of a wide range of process can be modeled by the GGD, such as DCT and wavelets coefficients and pixels difference [4,24]. The GGD pdf is given by

$$f(x) = \frac{k\alpha}{2\Gamma(1/k)} \exp -(\alpha|x - \theta|)^k \quad (3.6)$$

where $\Gamma(\cdot)$ is the gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, θ is the location parameter and α is a constant related to the standard deviation σ , defined as $\alpha = \sigma^{-1} \sqrt{\Gamma(3/k)(\Gamma(1/k))^{-1}}$. In this form, α is an inverse scale parameter and $k > 0$, sometimes called the shape parameter, controls the tail decay rate. The GGD model contains the Laplacian and Gaussian distributions as special cases, *i.e.*, for $k = 1$ and $k = 2$, respectively. Conceptually, the lower the value of k the more impulsive the distribution is. The ML location estimate for GGD statistics is reviewed in the following. Detailed derivations of these results are given in [4].

Consider a set of N independent observations each obeying the GGD with common location parameter, common shape parameter k and different scale parameter σ_i . The ML estimate of location is given by

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N \frac{1}{\sigma_i^k} |x(i) - \theta|^k \right]. \quad (3.7)$$

There are two special cases of the GGD family that are well studied: the Gaussian ($k = 2$) and the Laplacian ($k = 1$) distributions, which yield the well known *weighted mean* and *weighted median* estimators, respectively. When all samples are identically distributed for the special cases, the *mean* and *median* estimators are the resulting operators. These estimators are formally defined in the following.

Definition 3. Consider a set of N independent observations each obeying the Gaussian distribution with different variance σ_i^2 . The ML estimate of location is given by

$$\hat{\theta} = \frac{\sum_{i=1}^N h_i x(i)}{\sum_{i=1}^N h_i} \triangleq \text{mean}\{h_i \cdot x(i)|_{i=1}^N\} \quad (3.8)$$

where $h_i = 1/\sigma_i^2$ and \cdot denotes the (multiplicative) weighting operation.

Definition 4. Consider a set of N independent observations each obeying the Laplacian distribution with common location and different scale parameter σ_i . The ML

estimate of location is given by

$$\hat{\theta} = \text{median}\{h_i \diamond x(i)|_{i=1}^N\}. \quad (3.9)$$

where $h_i = 1/\sigma_i$ and \diamond denotes the replication operator defined as

$$h_i \diamond x(i) = \overbrace{x(i), x(i), \dots, x(i)}^{h_i \text{ times}}.$$

Through arguments similar to those above, the $k \neq 1, 2$ cases yield the fractional lower order moment (FLOM) estimation framework [154]. For $k < 1$, the resulting estimators are selection type. A drawback of FLOM estimators for $1 < k < 2$ is that their computation is, in general, nontrivial, although suboptimal (for $k > 1$) selection-type FLOM estimators have been introduced to reduce computational costs [6].

3.2.3 Generalized Cauchy Distribution

The GCD family was proposed by Rider in 1957 [150], rediscovered by Miller and Thomas in 1972 with a different parametrization [132], and has been used in several studies of impulsive radio noise [4, 7, 106, 107, 132]. The GCD pdf is given by

$$f_{GC}(z) = a\sigma(\sigma^p + |z - \theta|^p)^{-\frac{2}{p}} \quad (3.10)$$

with $a = p\Gamma(2/p)/2(\Gamma(1/p))^2$. In this representation, θ is the location parameter, σ is the scale parameter, and p is the tail constant. The GCD family contains the Meridian [9] and Cauchy distributions as special cases, *i.e.* for $p = 1$ and $p = 2$, respectively. For $p < 2$, the tail of the pdf decays slower than in the Cauchy distribution case, resulting in a heavier-tailed distribution.

The flexibility and closed-form nature of the GCD make it an ideal family

from which to derive robust estimation and filtering techniques. As such, we consider the location estimation problem that, as in the previous case, is approached from a ML estimation framework. Thus consider a set of N i.i.d. GCD distributed samples with common scale parameter σ and tail constant p . The ML estimate of location is given by

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N \log \{ \sigma^p + |x(i) - \theta|^p \} \right]. \quad (3.11)$$

Next, consider a set of N independent observations each obeying the GCD with common tail constant p , but possessing unique scale parameter ν_i . The ML estimate is formulated as $\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^N f_{GC}(x(i); \nu_i)$. Inserting the GCD distribution for each sample, taking the natural log and utilizing basic properties of the arg max and log functions yields

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \log \left[\prod_{i=1}^N a \nu_i (\nu_i^p + |x(i) - \theta|^p)^{-\frac{2}{p}} \right] \\ &= \arg \max_{\theta} \sum_{i=1}^N -\frac{2}{p} \log \{ \nu_i^p + |x(i) - \theta|^p \} \\ &= \arg \min_{\theta} \sum_{i=1}^N \log \left\{ 1 + \frac{|x(i) - \theta|^p}{\nu_i^p} \right\} \\ &= \arg \min_{\theta} \sum_{i=1}^N \log \{ \sigma^p + h_i |x(i) - \theta|^p \} \end{aligned} \quad (3.12)$$

with $h_i = (\sigma/\nu_i)^p$.

Since the estimator defined in (3.11) is a special case of that defined in (3.12), we only provide a detailed derivation for the latter. The estimator defined in (3.12) can be used to extend the GCD-based estimator to a robust weighted filter structure. Furthermore, the derived filter can be extended to admit real-valued weights using the sign-coupling approach [3].

3.2.4 Statistical Relationship Between the Generalized Cauchy and Gaussian Distributions

Before closing this section, we bring to light an interesting relationship between the Generalized Cauchy and Generalized Gaussian distributions. It is well-known that a Cauchy distributed random variable (GCD $p = 2$) is generated by the ratio of two independent Gaussian distributed random variables (GGD $k = 2$). Recently, Aysal and Barner showed that this relationship also holds for the Laplacian and Meridian distributions [9], *i.e.*, the ratio of two independent Laplacian (GGD $k = 1$) random variables yields a Meridian (GCD $p = 1$) random variable. In the following, we extend this finding to the complete set of GGD and GCD families.

Lemma 1. *The random variable formed as the ratio of two independent zero-mean GGD distributed random variables U and V , with tail constant β and scale parameters α_U and α_V , respectively, is a GCD random variable with tail parameter $\lambda = \beta$ and scale parameter $\nu = \alpha_U/\alpha_V$.*

Proof. See Appendix A. □

3.3 Generalized Cauchy Based Robust Estimation and Filtering

In this section we use the GCD ML location estimate cost function to define an M-type estimator. First, robustness and properties of the derived estimator are analyzed and the filtering problem is then related to M-estimation. The proposed estimator is extended to a weighted filtering structure. Finally, practical algorithms for the multi-parameter case are developed.

3.3.1 Generalized Cauchy Based M-Estimation

The cost function associated with the GCD ML estimate of location derived in the previous section is given by

$$\rho(x) = \log\{\sigma^p + |x|^p\}, \quad \sigma > 0, \quad 0 < p \leq 2. \quad (3.13)$$

The flexibility of this cost function, provided by parameters σ and p , and robust characteristics make it well-suited to define an M-type estimator, which we coin the *M-GC* estimator. To define the form of this estimator, denote $\mathbf{x} = [x(1), \dots, x(N)]$ as a vector of observations and θ the common location parameter of the observations.

Definition 5. *The M-GC estimate is defined as*

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N \log\{\sigma^p + |x(i) - \theta|^p\} \right]. \quad (3.14)$$

The special $p = 2$ and $p = 1$ cases yield the *myriad* [108] and *meridian* [9] estimators, respectively. The generalization of the M-GC estimator, for $0 < p \leq 2$, is analogous to the GGD-based FLOM estimators and thereby provides a rich and robust framework for signal processing applications.

As the performance of an estimator depends on the defining objective function, the properties of the objective function at hand are analyzed in the following.

Proposition 1. *Let $Q(\theta) = \sum_{i=1}^N \log\{\sigma^p + |x(i) - \theta|^p\}$ denote the objective function (for fixed σ and p) and $\{x_{[i]}\}_{i=1}^N$ the order statistics of \mathbf{x} . Then the following statements hold.*

1. $Q(\theta)$ is strictly decreasing for $\theta < x_{[1]}$ and strictly increasing for $\theta > x_{[N]}$.
2. All local extrema of $Q(\theta)$ lie in the interval $[x_{[1]}, x_{[N]}]$.
3. If $0 < p \leq 1$, the solution is one of the input samples (selection type filter).
4. If $1 < p \leq 2$, then the objective function has at most $2N - 1$ local extrema points and therefore a finite set of local minima.

Proof. See Appendix B. □

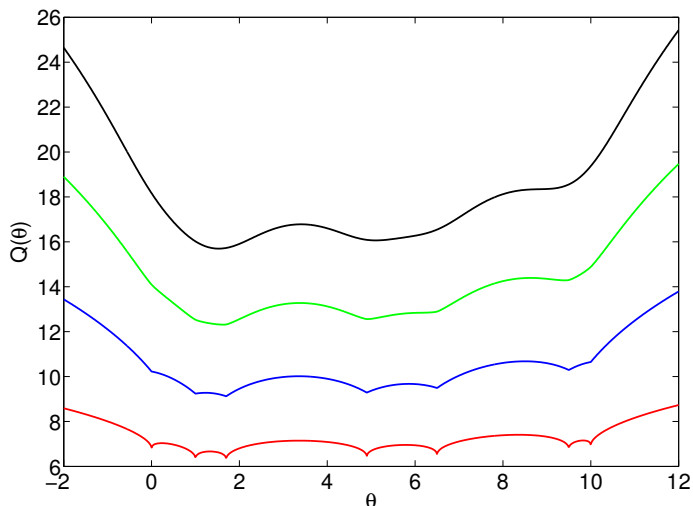


Figure 3.1: Typical M-GC objective functions for different values of $p \in \{0.5, 1, 1.5, 2\}$ (from bottom to top respectively). Input samples are $\mathbf{x} = [4.9, 0, 6.5, 10.0, 9.5, 1.7, 1]$ and $\sigma = 1$.

The M-GC estimator has two adjustable parameters, σ and p . The tail constant, p , depends on the heaviness of the underlying distribution. Notably, when $p \leq 1$ the estimator behaves as a selection type filter and, as $p \rightarrow 0$, it becomes increasingly robust to outlier samples. For $p > 1$, the location estimate is in the range of the input samples and is readily computed. Fig. 3.1 shows a typical sketch of the M-GC objective function, in this case for $p \in \{0.5, 1, 1.5, 2\}$ and $\sigma = 1$.

The following properties detail the M-GC estimator behavior as σ goes to either 0 or ∞ . Importantly, the results show that the M-GC estimator subsumes other classical estimator families.

Property 1. *Given a set of input samples $\{x(i)\}_{i=1}^N$, the M-GC estimate converges to the ML GGD estimate (L_p norm as cost function) as $\sigma \rightarrow \infty$.*

$$\lim_{\sigma \rightarrow \infty} \hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N |x(i) - \theta|^p. \quad (3.15)$$

Proof. Using the properties of the arg min function the M-GC estimator can be expressed as

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \log \left\{ 1 + \frac{|x(i) - \theta|^p}{\sigma^p} \right\}. \quad (3.16)$$

Let $\delta = \sigma^p$. Since multiplying by a constant does not affect the result of the arg min operator we can rewrite (3.16) as

$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^N \delta \log \left\{ 1 + \frac{|x(i) - \theta|^p}{\delta} \right\}.$$

Using the fact that $a \log b = \log b^a$ and taking the limit as $\delta \rightarrow \infty$ yields

$$\begin{aligned} \lim_{\delta \rightarrow \infty} \hat{\theta} &= \lim_{\delta \rightarrow \infty} \arg \min_{\theta} \sum_{i=1}^N \log \left\{ 1 + \frac{|x(i) - \theta|^p}{\delta} \right\}^{\delta} \\ &= \arg \min_{\theta} \sum_{i=1}^N |x(i) - \theta|^p, \end{aligned} \quad (3.17)$$

where the last step follows since

$$\lim_{\delta \rightarrow \infty} \log \left\{ 1 + \frac{u}{\delta} \right\}^{\delta} = u. \quad \square$$

Intuitively, this result is explained by the fact that $|x(i) - \theta|^p/\sigma^p$ becomes negligible as σ grows large compared to 1. This, combined with the fact that $\log(1 + x) \approx x$ when $x \ll 1$, which is an equality in the limit, yields the resulting cost function behavior. The importance of this result is that M-GC estimators include M-estimators with L_p norm ($0 < p \leq 2$) cost functions. Thus M-GC (GCD-based) estimators should be at least as powerful as GGD-based estimators (linear FIR, median, FLOM) in light-tailed applications, while the untapped algebraic tail potential of GCD methods should allow them to substantially outperform in heavy-tailed applications.

In contrast to the equivalence with L_p norm approaches for σ large, M-GC estimators becomes more resistant to impulsive noise as σ decreases. In fact, as $\sigma \rightarrow 0$ the M-GC yields a mode type estimator with particularly strong impulse rejection.

Property 2. *Given a set of input samples $\{x(i)\}_{i=1}^N$, the M-GC estimate converges to a mode type estimator as $\sigma \rightarrow 0$. This is*

$$\lim_{\sigma \rightarrow 0} \hat{\theta} = \arg \min_{x(j) \in \mathcal{M}} \left[\prod_{i, x(i) \neq x(j)} |x(i) - x(j)| \right] \quad (3.18)$$

where \mathcal{M} is the set of most repeated values.

Proof. The M-GC estimator can be expressed as

$$\begin{aligned} \hat{\theta} &= \arg \min_{\theta} \sum_{i=1}^N \log \left\{ 1 + \frac{|x(i) - \theta|^p}{\sigma^p} \right\} \\ &= \arg \min_{\theta} \log \left\{ \prod_{i=1}^N \left[1 + \frac{|x(i) - \theta|^p}{\sigma^p} \right] \right\}. \end{aligned} \quad (3.19)$$

Define

$$H_{\sigma}(\theta; \mathbf{x}) = \prod_{i=1}^N \left[1 + \frac{|x(i) - \theta|^p}{\sigma^p} \right]. \quad (3.20)$$

Since the log function is monotone nondecreasing the M-GC estimator can be reformulated as:

$$\hat{\theta} = \arg \min_{\theta} H_{\sigma}(\theta; \mathbf{x}).$$

It can be checked that when σ is very small

$$H_{\sigma}(\theta; \mathbf{x}) = \mathcal{O} \left(\frac{1}{\sigma^p} \right)^{N-r(\theta)} \quad (3.21)$$

where $r(\theta)$ is the number of times the value θ is repeated in the sample set and \mathcal{O} denotes the asymptotic order as $\sigma \rightarrow 0$. In the limit the exponent $N - r(\theta)$

must be minimized for $H_\sigma(\theta; \mathbf{x})$ to be minimum. Therefore, $\hat{\theta}$ will be one of the most repeated values in the input set. Define $r = \max_j r(x(j))$, then for $x(j) \in \mathcal{M}$, expanding the product in (3.20) gives

$$H_\sigma(x(j); \mathbf{x}) = \left\{ \prod_{i, x(i) \neq x(j)} \frac{|x(i) - \theta|^p}{\sigma^p} \right\} + \mathcal{O}\left(\frac{1}{\sigma^p}\right)^{N-r-1}. \quad (3.22)$$

Since the first term in (3.22) is $\mathcal{O}(1/\sigma^p)^{N-r}$, the second term is negligible for small σ . Then, in the limit, $\hat{\theta}$ can be computed as

$$\begin{aligned} \lim_{\sigma \rightarrow 0} \hat{\theta} &= \arg \min_{x(j) \in \mathcal{M}} [H_\sigma(x(j); \mathbf{x})] \\ &= \arg \min_{x(j) \in \mathcal{M}} \left[\prod_{i, x(i) \neq x(j)} \frac{|x(i) - x(j)|^p}{\sigma^p} \right] \\ &= \arg \min_{x(j) \in \mathcal{M}} \left[\prod_{i, x(i) \neq x(j)} |x(i) - x(j)| \right]. \quad \square \end{aligned} \quad (3.23)$$

This mode-type estimator treats every observation as a possible outlier, assigning greater influence to the most repeated values in the observations set. This property makes the M-GC a suitable framework for applications such as image processing, where selection-type filters yield good results [9, 108, 170].

3.3.2 Robustness and Analysis of M-GC Estimators

To formally evaluate the robustness of M-GC estimators we consider the influence function, which, if it exists, is proportional to $\psi(x)$ and determines the effect of contamination of the estimator. For the M-GC estimator

$$\psi(x) = \frac{p|x|^{p-1} \text{sgn}(x)}{\sigma^p + |x|^p} \quad (3.24)$$

where $\text{sgn}(\cdot)$ denotes the sign operator. Fig. 3.2 shows the M-GC estimator influence function for $p \in \{0.5, 1, 1.5, 2\}$.

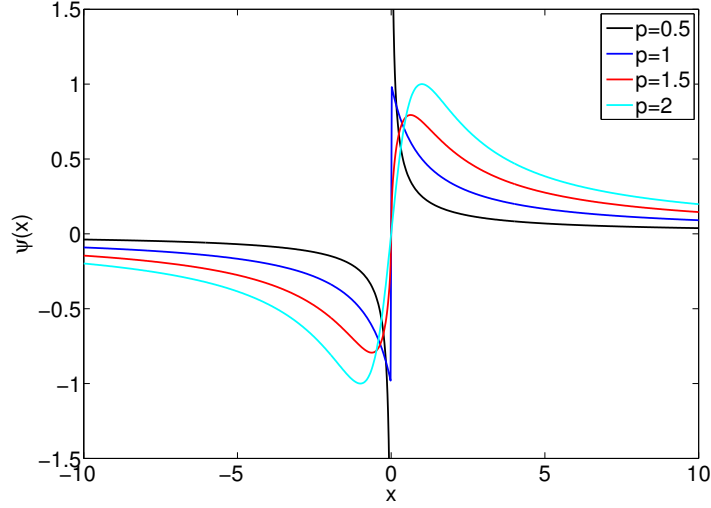


Figure 3.2: Influence functions of the M-GC estimator for different values of p . (black:) $p = 0.5$, (blue:) $p = 1$, (red:) $p = 1.5$, and (cyan:) $p = 2$.

To further characterize M-estimates, it is useful to list the desirable features of a robust influence function [113, 118].

- *B-robustness:* An estimator is B-robust if the supremum of the absolute value of the influence function is finite.
- *Rejection Point:* The rejection point, defined as the distance from the center of the influence function to the point where the influence function becomes negligible, should be finite. Rejection point measures whether the estimator rejects outliers and, if so, at what distance.

The M-GC estimate is B-robust and has a finite rejection point that depends on the scale parameter σ and the tail parameter p . As $p \rightarrow 0$, the influence function has higher decay rate, *i.e.*, as $p \rightarrow 0$ the M-GC estimator becomes more robust to outliers. Also of note is that $\lim_{x \rightarrow \pm\infty} \psi(x) = 0$, *i.e.*, the influence function is asymptotically re-descending and the effect of outliers monotonically decreases with an increase in magnitude [113].

The M-GC also possesses the followings important properties.

Property 3. (*Outlier Rejection*) For $\sigma < \infty$,

$$\lim_{x(N) \rightarrow \pm\infty} \hat{\theta}(x(1), \dots, x(N)) = \hat{\theta}(x(1), \dots, x(N-1)). \quad (3.25)$$

Property 4. (*No undershoot/overshoot*) The output of the M-GC estimator is always bounded by

$$x_{[1]} < \hat{\theta} < x_{[N]} \quad (3.26)$$

where $x_{[1]} = \min\{x(i)\}_{i=1}^N$ and $x_{[N]} = \max\{x(i)\}_{i=1}^N$.

According to Property 3, large errors are efficiently eliminated by an M-GC estimator with finite σ . Note that this property can be applied recursively, indicating M-GC estimators eliminate multiple outliers. The proof of this statement follows the same steps used in the proof of the meridian estimator Property 9 [9], and is thus omitted. Property 4 states that the M-GC estimator is BIBO stable, *i.e.*, the output is bounded for bounded inputs. Proof of Property 4 follows directly from Proposition 1-2 and is thus omitted.

Since M-GC estimates are M-estimates, they have desirable asymptotic behavior, as noted in the following Theorem and discussion.

Theorem 5. (*Asymptotic Consistency*) Suppose the samples $\{x(i)\}_{i=1}^N$ are independent and symmetrically distributed around θ (location parameter). Then, the M-GC estimate $\hat{\theta}_N$, converges to θ in probability, *i.e.*,

$$\hat{\theta}_N \xrightarrow{P} \theta \text{ as } N \rightarrow \infty. \quad (3.27)$$

Proof. The proof follows from the fact that the M-GC estimator influence function is odd, bounded, and continuous (except at the origin, which is a set of measure

zero); argument details parallel those in [118]. Define

$$\lambda(s) = \mathbb{E}_F \psi(X - s) \quad (3.28)$$

and

$$\hat{\lambda}_N(s) = \frac{1}{N} \sum_{i=1}^N \psi(x(i) - s), \quad (3.29)$$

where the expectation is taken with respect to F , the underlying distribution of X , and $\psi(x)$ is the influence function of the M-GC estimator. From the definition of $\psi(x)$ in equation (3.24) we can see that

- $\psi(x)$ is an odd function in x and therefore $\psi(x - s)$ is odd in s .
- $\psi(x - s) > 0$ if $x - s < 0$ and $\psi(x - s) < 0$ if $x - s > 0$.

It can be noticed that $\{\psi(x(i) - s)\}_{i=1}^N$ are i.i.d. random variables with finite variance for any s . Then by the weak law of large numbers the following holds

$$\hat{\lambda}_N(s) \xrightarrow{P} \lambda(s) \text{ as } N \rightarrow \infty. \quad (3.30)$$

It can be shown that if θ is the location parameter of X then $\lambda(\theta) = 0$, since ψ is odd and F is symmetric around θ . From the definition of the M-GC estimate we have that $\hat{\lambda}_N(\hat{\theta}_N) = 0$.

Let ϵ be a positive constant. Thus, $\lambda(\theta - \epsilon) > 0$ and $\hat{\lambda}_N(\theta - \epsilon) \xrightarrow{P} \lambda(\theta - \epsilon)$. The above implies that $\lim_{N \rightarrow \infty} P(\hat{\lambda}_N(\theta - \epsilon) > 0) = 1$. Since $\hat{\lambda}_N(\theta - \epsilon) > 0$ implies $\theta - \epsilon < \hat{\theta}_N$ and viceversa, it follows that $\lim_{N \rightarrow \infty} P(\theta - \hat{\theta}_N < \epsilon) = 1$.

Similarly, it can be shown that $\lim_{N \rightarrow \infty} P(\theta - \hat{\theta}_N > -\epsilon) = 1$. Therefore $\hat{\theta}_N \xrightarrow{P} \theta$ as $N \rightarrow \infty$ since ϵ is an arbitrary positive constant. \square

Notably, M-estimators have asymptotic normal behavior [118]. In fact, it can be shown that

$$\sqrt{N}(\hat{\theta}_N - \theta) \rightarrow Z \quad (3.31)$$

in distribution, where $Z \sim \mathcal{N}(0, v)$ and

$$v = \frac{E_F \psi^2(X - \theta)}{(E_F \psi'(X - \theta))^2}. \quad (3.32)$$

The expectation is taken with respect to F , the underlying distribution of the data. The last expression is the asymptotic variance of the estimator. Hence, the variance of $\hat{\theta}_N$ decreases as N increases, meaning that M-GC estimates are asymptotically efficient.

3.3.3 Weighted M-GC Estimators

A filtering framework cannot be considered complete until an appropriate weighting operation is defined. Filter weights, or coefficients, are extremely important for applications in which signal correlations are to be exploited. Using the ML estimator under independent, but non identically distributed, GCD statistics (expression (3.12)), the M-GC estimator is extended to include weights. Let $\mathbf{h} = [h_1, \dots, h_N]$ denote a vector of non-negative weights. The weighted M-GC (WM-GC) estimate is defined as

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N \log\{\sigma^p + h_i |x(i) - \theta|^p\} \right]. \quad (3.33)$$

The filtering structure defined in (3.33) is an M-smoother estimator, which is in essence a low-pass-type filter. Utilizing the sign coupling technique [3], the M-GC estimator can be extended to accept real-valued weights. This yields the general structure detailed in the following definition.

Definition 6. *The weighted M-GC (WM-GC) estimate is defined as*

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_{i=1}^N \log \{ \sigma^p + |h_i| | \text{sgn}(h_i)x(i) - \theta|^p \} \right] \quad (3.34)$$

where $\mathbf{h} = [h_1, \dots, h_N]$ denotes a vector of real valued weights.

The WM-GC estimators inherit all the robustness and convergence properties of the unweighted M-GC estimators. Thus as in the unweighted case, WM-GC estimators subsume GGD-based (weighted) estimators, indicating that WM-GC estimators are at least as powerful as GGD-based estimators (linear FIR, weighted median, weighted FLOM) in light-tailed environments, while WM-GC estimator characteristics enable them to substantially outperform in heavy-tailed impulsive environments.

3.3.4 Multi-parameter Estimation

The location estimation problem defined by the M-GC filter depends on the parameters σ and p . Thus to solve the optimal filtering problem, we consider multi-parameter M-estimates [48]. The applied approach utilizes a small set of signal samples to estimate σ and p , and then uses these values in the filtering process (although a fully adaptive filter can also be implemented using this scheme).

Let $\{x(i)\}_{i=1}^N$ be a set of independent observations from a common GCD with deterministic but unknown parameters θ , σ and p . The joint estimates are the solutions to the following maximization problem

$$(\hat{\theta}, \hat{\sigma}, \hat{p}) = \arg \max_{\theta, \sigma, p} g(\mathbf{x}; \theta, \sigma, p) \quad (3.35)$$

where

$$g(\mathbf{x}; \theta, \sigma, p) = \prod_{i=1}^N a \sigma (\sigma^p + |x(i) - \theta|^p)^{-\frac{2}{p}}, \quad a = p\Gamma(2/p)/2(\Gamma(1/p))^2. \quad (3.36)$$

The solution to this optimization problem is obtained by solving a set of simultaneous equations given by first order optimality conditions. Differentiating the log-likelihood function, $g(\mathbf{x}; \theta, \sigma, p)$, with respect to θ , σ and p and performing some algebraic manipulations yields the following set of simultaneous equations:

$$\frac{\partial g}{\partial \theta} = \sum_{i=1}^N \frac{-p|x(i) - \theta|^{p-1} \text{sgn}(x(i) - \theta)}{\sigma^p + |x(i) - \theta|^p} = 0 \quad (3.37)$$

$$\frac{\partial g}{\partial \sigma} = \sum_{i=1}^N \frac{\sigma^p - |x(i) - \theta|^p}{\sigma^p + |x(i) - \theta|^p} = 0 \quad (3.38)$$

and

$$\begin{aligned} \frac{\partial g}{\partial p} = \sum_{i=1}^N \left[\frac{1}{2p} - \frac{\sigma^p \log \sigma - |x(i) - \theta|^p \log |x(i) - \theta|}{p(\sigma^p - |x(i) - \theta|^p)} \right. \\ \left. - \frac{\log\{\sigma^p + |x(i) - \theta|^p\}}{p^2} - \frac{1}{p^2} \Psi\left(\frac{2}{p}\right) + \frac{1}{p^2} \Psi\left(\frac{1}{p}\right) \right] = 0. \end{aligned} \quad (3.39)$$

where $g \equiv g(\mathbf{x}; \theta, \sigma, p)$ and $\Psi(x)$ is the digamma function². It can be noticed that (3.37) is the implicit equation for the M-GC estimator with ψ as defined in (3.24), implying that the location estimate has the same properties derived above.

Of note is that $g(\mathbf{x}; \theta, \sigma, p)$ has a unique maximum in σ for fixed θ and p , and also a unique maximum in p for fixed θ and σ and $p \in (0, 2]$. In the following, we provide an algorithm to iteratively solve the above set of equations.

Multi-parameter Estimation Algorithm: For a given set of data $\{x(i)\}_{i=1}^N$, we propose to find the optimal joint parameter estimates by the iterative algorithm details in Algorithm 3, with the superscript denoting iteration number.

The algorithm is essentially an iterated conditional mode (ICM) algorithm [29]. Additionally, it resembles the expectation maximization (EM) algorithm [129] in the

² The digamma function is defined as $\Psi(x) = \frac{d}{dx} \Gamma(x)$, where $\Gamma(x)$ is the Gamma function.

Algorithm 3 Multi-parameter Estimation Algorithm

Require: Data set $\{x(i)\}_{i=1}^N$ and tolerances $\epsilon_1, \epsilon_2, \epsilon_3$.

- 1: Initialize $\sigma^{(0)}$ and $\theta^{(0)}$.
 - 2: **while** $|\hat{\theta}^{(m)} - \hat{\theta}^{(m-1)}| > \epsilon_1$, $|\hat{\sigma}^{(m)} - \hat{\sigma}^{(m-1)}| > \epsilon_2$ and $|\hat{p}^{(m)} - \hat{p}^{(m-1)}| > \epsilon_3$ **do**
 - 3: Estimate $\hat{p}^{(m)}$ as the solution of (3.39).
 - 4: Estimate $\hat{\theta}^{(m)}$ as the solution of (3.37).
 - 5: Estimate $\hat{\sigma}^{(m)}$ as the solution of (3.38).
 - 6: **end while**
 - 7: **return** $\hat{\theta}, \hat{\sigma}$ and \hat{p} .
-

sense that, instead of optimizing all parameters at once, it finds the optimal value of one parameter given that the other two are fixed; it then iterates. While the algorithm converges to a local minimum, experimental results show that initializing θ as the sample median and σ as the median absolute deviation (MAD), and then computing p as a solution to (3.39), accelerates the convergence and most often yields globally optimal results. In the classical literature fixed point algorithms are successfully used in the computation of M-estimates [4, 118]. Hence, in the following, we solve items 3-5 in Algorithm 3 using fixed point search routines.

Fixed-Point Search Algorithms: Recall that when $0 < p \leq 1$, the solution is the input sample that minimizes the objective function. We solve (3.37) for the $1 < p \leq 2$ case using the fixed point recursion, which can be written as

$$\hat{\theta}_{(j+1)} = \frac{\sum_{i=1}^N w_i(\hat{\theta}_{(j)})x(i)}{\sum_{i=1}^N w_i(\hat{\theta}_{(j)})} \quad (3.40)$$

with $w_i(\hat{\theta}_{(j)}) = p|x(i) - \hat{\theta}_{(j)}|^{p-2}/(\sigma^p + |x(i) - \hat{\theta}_{(j)}|^p)$ and where the subscript denotes the iteration number. The algorithm is taken as convergent when $|\hat{\theta}_{(j+1)} - \hat{\theta}_{(j)}| < \delta_1$, where δ_1 is a small positive value. The median is used as the initial estimate, which typically results in convergence to a (local) minima within a few iterations.

Similarly, for (3.38) the recursion can be written as

$$\hat{\sigma}_{(j+1)} = \left(\frac{\sum_{i=1}^N b_i(\hat{\sigma}_{(j)})x(i)}{\sum_{i=1}^N b_i(\hat{\sigma}_{(j)})} \right)^{\frac{1}{p}} \quad (3.41)$$

with $b_i(\hat{\sigma}_{(j)}) = 1/(\hat{\sigma}_{(j)}^p + |x(i) - \theta|^p)$. The algorithm terminates when $|\hat{\sigma}_{(j+1)} - \hat{\sigma}_{(j)}| < \delta_2$ for δ_2 a small positive number. Since the objective function has only one minimum for fixed θ and p , the recursion converges to the global result.

The parameter p recursion is given by

$$\begin{aligned} \hat{p}_{(j+1)} = & \frac{2}{N} \sum_{i=1}^N \left[\Psi\left(\frac{2}{\hat{p}_{(j)}}\right) - \Psi\left(\frac{1}{\hat{p}_{(j)}}\right) + \log\{\sigma^{\hat{p}_{(j)}} + |x(i) - \theta|^{\hat{p}_{(j)}}\} \right. \\ & \left. + \frac{\hat{p}_{(j)}(\sigma^{\hat{p}_{(j)}} \log \sigma - |x(i) - \theta|^{\hat{p}_{(j)}} \log |x(i) - \theta|)}{\sigma^{\hat{p}_{(j)}} - |x(i) - \theta|^{\hat{p}_{(j)}}} \right]. \end{aligned} \quad (3.42)$$

Noting that the search space is the interval $I = (0, 2]$, the function g (equation (3.36)) can be evaluated for a finite set of points $\mathcal{P} \in I$, keeping the value that maximizes g , setting it as the initial point for the search.

As an example, simulations illustrating the developed multi-parameter estimation algorithm are summarized in Table 3.1, for $p = 2$, $\theta = 0$ and $\sigma = 1$ (standard Cauchy distribution). Results are shown for varying sample lengths; 10, 100, and 1000. The experiments were run 1000 times for each block length, with the presented results the average on the trials. Mean final θ , σ , and p estimates are reported as well as the resulting MSE. To illustrate that the algorithm converges in a few iterations, given the proposed initialization, consider an an experiment utilizing data drawn from a GCD $\theta = 0$, $\sigma = 1$ and $p = 1.5$ distribution. Fig. 3.3 reports θ , σ , p estimate MSE curves. As in the previous case, 100 trials are averaged. Only the first five iteration points are shown, as the algorithms are convergent at that point.

To conclude this section, we consider the computational complexity of the proposed multi-parameter estimation algorithm. The algorithm in total has a higher

Table 3.1: Multi-parameter Estimation Results for GCD Process with length N and $(\theta, \sigma, p) = (0, 1, 2)$.

N	10	100	1000
$\hat{\theta}$	0.0035	-0.0009	-0.0002
MSE	0.0302	2.4889×10^{-3}	1.7812×10^{-4}
$\hat{\sigma}$	0.9563	1.0224	1.0186
MSE	0.0016	1.7663×10^{-5}	1.1911×10^{-6}
\hat{p}	1.5816	1.8273	1.9569
MSE	0.0519	0.0109	1.5783×10^{-6}

computational complexity than the FLOM, median, meridian, and myriad operators, since Algorithm 3 requires initial estimates of the location and the scale parameters. However, it should be noted that the proposed method estimates *all* the parameters of the model, thus providing advantage over the aforementioned methods that require *a priori* parameter tuning. It is straightforward to show that the computational complexity of the proposed method is $\mathcal{O}(N^2)$, assuming the practical case in which the number of fixed point iterations is $\ll N$. The dominating N^2 term is the cost of selecting the input sample that minimizes the objective function, *i.e.*, the cost of evaluating the objective function N times. However, if faster methods that avoid evaluation of the objective function for all samples (*e.g.*, subsampling methods) are employed, the computational cost is lowered.

3.4 Robust Distance Metrics

This section presents a family of robust GCD based error metrics. Specifically, the cost function of the M-GC estimator defined in Section 3.3.1 is extended to define a quasi-norm over \mathbb{R}^m and a semimetric for the same space – the development is analogous to L_p norms emanating from the GGD family. We denote these semimetrics as the log- L_p (LL_p) norms³.

³ Note that for the $\sigma = 1$ and $p = 1$ case, this metric defines the log- L space in Banach space theory.

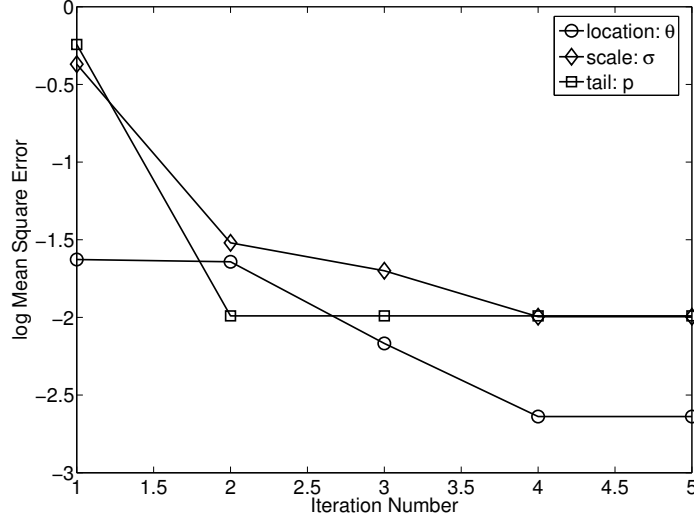


Figure 3.3: Multi-parameter estimation MSE iteration evolution for a GCD process with $(\theta, \sigma, p) = (0, 1, 1.5)$.

Definition 7. Let $u \in \mathbb{R}^m$, then the LL_p norm of u is defined as

$$\|u\|_{LL_p, \sigma} = \sum_{i=1}^m \log \left\{ 1 + \frac{|u_i|^p}{\sigma^p} \right\}, \quad \sigma > 0. \quad (3.43)$$

The LL_p norm is not a norm in the strictest sense since it does not meet the positive homogeneity and subadditivity properties. However, it follows the positive definiteness and a scale invariant properties.

Proposition 2. Let $c \in \mathbb{R}$, $u, v \in \mathbb{R}^m$, and $p, \sigma > 0$. The following statements hold

- i) $\|u\|_{LL_p, \sigma} \geq 0$, with $\|u\|_{LL_p, \sigma} = 0$ if and only if $u = 0$.
- ii) $\|cu\|_{LL_p, \sigma} = \|u\|_{LL_p, \delta}$ where $\delta = \sigma/|c|$.
- iii) $\|u + v\|_{LL_p, \sigma} = \|v + u\|_{LL_p, \sigma}$.
- iv) $\|u + v\|_{LL_p, \sigma} \leq \begin{cases} \|u\|_{LL_p, \sigma} + \|v\|_{LL_p, \sigma}, & \text{for } 0 < p \leq 1, \\ \|u\|_{LL_p, \sigma} + \|v\|_{LL_p, \sigma} + m \log C_p, & \text{for } p > 1, \end{cases}$, where $C_p = 2^{p-1}$.

Proof. Statement 1 follows from the fact that $\log(1 + a) \geq 0$ for all $a \geq 0$, with equality if and only if $a = 0$. Statement 2 follows from

$$\sum_{i=1}^m \log \left\{ 1 + \frac{|cu_i|^p}{\sigma^p} \right\} = \sum_{i=1}^m \log \left\{ 1 + \frac{|u_i|^p}{(\sigma/|c|)^p} \right\}.$$

Statement 3 follows directly from the definition of the LL_p norm. Statement 4 follows from the well known relation: $|a + b|^p \leq C_p(|a|^p + |b|^p)$, $a, b \in \mathbb{R}$, where C_p is a constant that depends only on p . Indeed, for $0 < p \leq 1$ we have $C_p = 1$, whereas for $p > 1$ we have $C_p = 2^{p-1}$ (for further details see [114] for example). Using this result and properties of the \log function we have

$$\begin{aligned} \|u + v\|_{LL_{p,\sigma}} &= \sum_{i=1}^m \log \left\{ 1 + \frac{|u_i + v_i|^p}{\sigma^p} \right\} \\ &\leq \sum_{i=1}^m \log \left\{ 1 + \frac{C_p(|u_i|^p + |v_i|^p)}{\sigma^p} \right\} \\ &= \sum_{i=1}^m \log C_p + \log \left\{ \frac{1}{C_p} + \frac{(|u_i|^p + |v_i|^p)}{\sigma^p} \right\} \\ &\leq \sum_{i=1}^m \log C_p + \log \left\{ 1 + \frac{(|u_i|^p + |v_i|^p)}{\sigma^p} \right\} \\ &\leq \sum_{i=1}^m \log \left\{ 1 + \frac{|u_i|^p}{\sigma^p} + \frac{|v_i|^p}{\sigma^p} + \frac{|u_i|^p |v_i|^p}{\sigma^{2p}} \right\} + m \log C_p \\ &= \sum_{i=1}^m \log \left\{ \left(1 + \frac{|u_i|^p}{\sigma^p} \right) \left(1 + \frac{|v_i|^p}{\sigma^p} \right) \right\} + m \log C_p \\ &= \|u\|_{LL_{p,\sigma}} + \|v\|_{LL_{p,\sigma}} + m \log C_p. \quad \square \end{aligned}$$

The LL_p norm defines a robust metric that does not heavily penalize large deviations, with the robustness depending on the scale parameter σ and the exponent p . The following Lemma constructs a relationship between the L_p norms and the LL_p norms.

Lemma 2. For every $u \in \mathbb{R}^m$, $0 < p \leq 2$ and $\sigma > 0$ the following relations hold:

$$\sigma^p \|u\|_{LL_p, \sigma} \leq \|u\|_p^p \leq \sigma^p m (e^{\|u\|_{LL_p, \sigma}} - 1). \quad (3.44)$$

Proof. The first inequality comes from the relation $\log(1+x) \leq x, \forall x \geq 0$. Setting $x_i = |u_i|^p / \sigma^p$ and summing over i yields the result. The second inequality follows from

$$\begin{aligned} \|u\|_{LL_p, \sigma} &= \sum_{i=1}^m \log \left\{ 1 + \frac{|u_i|^p}{\sigma^p} \right\} \\ &\geq \max_i \log \left\{ 1 + \frac{|u_i|^p}{\sigma^p} \right\} = \log \left\{ 1 + \frac{\|u\|_\infty^p}{\sigma^p} \right\}. \end{aligned}$$

Noting $\|u\|_\infty \leq \sigma (e^{\|u\|_{LL_p, \sigma}} - 1)^{1/p}$ and $\|u\|_p^p \leq m \|u\|_\infty^p$ for all $p > 0$ gives the desired result. \square

The particular case $p = 2$ yields the well-known Lorentzian norm. The Lorentzian norm has desirable robust error metric properties:

- It is an everywhere continuous function.
- It is convex near the origin ($0 \leq u \leq \sigma$), behaving similar to an L_2 cost function for small variations.
- Large deviations are not heavily penalized as in the L_1 or L_2 norm cases, leading to a more robust error metric when the deviations contain gross errors.

Contour plots of select norms are shown in Fig. 3.4 for the two dimension case. Fig. 3.4 (a) and (c) show the L_2 and L_1 norms, respectively, while the LL_2 (Lorentzian) and LL_1 norms (for $\sigma = 1$) are shown in Figs. 3.4 (b) and (d), respectively. It can be seen from Fig. 3.4 (b) that the Lorentzian norm tends to behave like the L_2 norm for points within the unitary L_2 ball. Conversely, it gives

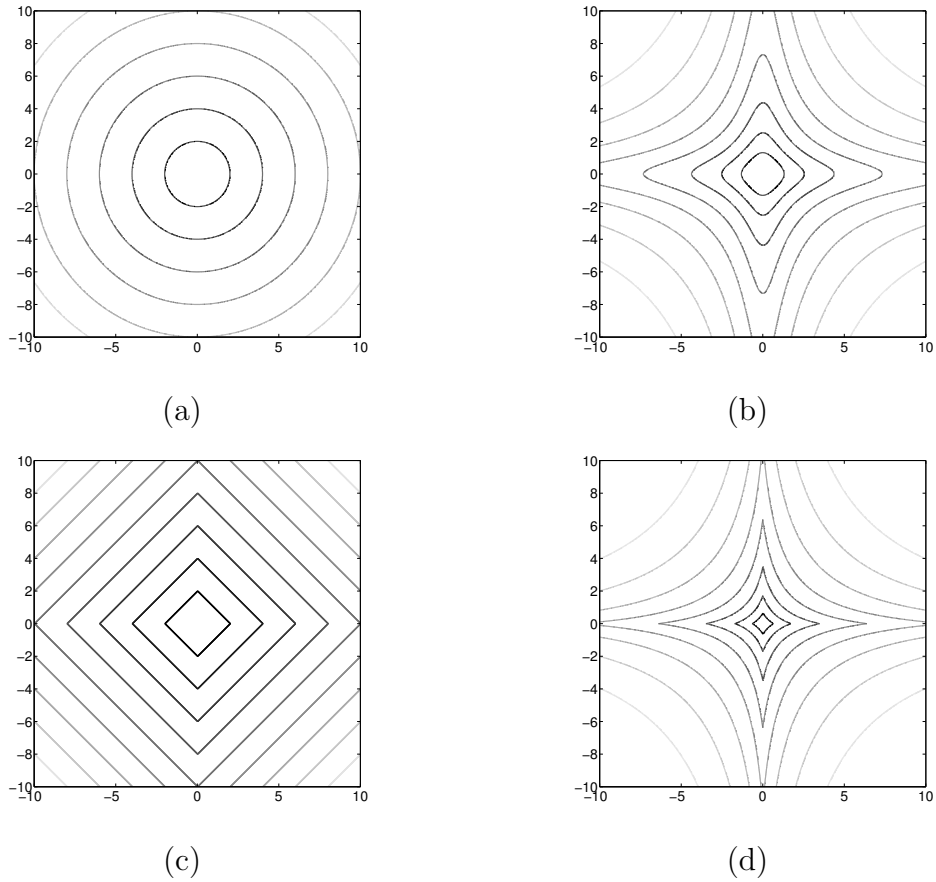


Figure 3.4: Contour plots of different metrics for two dimensions: (a) L_2 , (b) LL_2 (Lorentzian), (c) L_1 , and (d) LL_1 norms.

the same penalization to large sparse deviations as to smaller clustered deviations. In a similar fashion, Fig. 3.4 (d) shows that the LL_1 norm behaves like the L_1 norm for points in the unitary L_1 ball.

3.5 Illustrative Application Areas

This section presents three practical problems developed under the proposed framework: 1) robust filtering for power line communications, 2) robust estimation in sensor networks with noisy channels and 3) robust fuzzy clustering. Each problem serves to illustrate the capabilities and performance of the proposed methods.

3.5.1 Robust Filtering

The use of existing power lines for transmitting data and voice has been receiving recent interest [127, 171]. The advantages of power line communications (PLCs) are obvious due to the ubiquity of power lines and power outlets. The potential of power lines to deliver broadband services, such as fast internet access, telephone, fax services, and home networking is emerging in new communications industry technology. However, there remain considerable challenges for PLCs, such as communications channels that are hampered by the presence of large amplitude noise superimposed on top of traditional white Gaussian noise. The overall interference is appropriately modeled as an algebraic tailed process, with α -stable often chosen as the parent distribution [127].

While the M-GC filter is optimal for GCD noise, is also robust in general impulsive environments. To compare the robustness of the M-GC filter with other robust filtering schemes, experiments for symmetric α -stable noise corrupted PLCs are presented. Specifically, signal enhancement for the power line communication problem with a 4-ASK signaling, and equiprobable alphabet $v = \{-2, -1, 1, 2\}$, is considered. The noise is taken to be white, zero location, α -stable distributed with $\gamma = 1$ and α ranging from 0.2 to 2 (very impulsive to Gaussian noise). The filtering process employed utilize length nine sliding windows to remove the noise and enhance the signal. The M-GC parameters were determined using the multi-parameter estimation algorithm described in Section 3.3.4. This optimization was applied to the first 50 samples, yielding $p = 0.756$ and $\sigma = 0.896$. The M-GC filter is compared to the FLOM, median, myriad, and meridian operators. The meridian tunable parameter was also set using the multi-parameter optimization procedure, but without estimating p . The myriad filter tuning parameter was set according to the $\alpha - k$ curve established in [108].

The normalized MSE values for the outputs of the different filtering structures

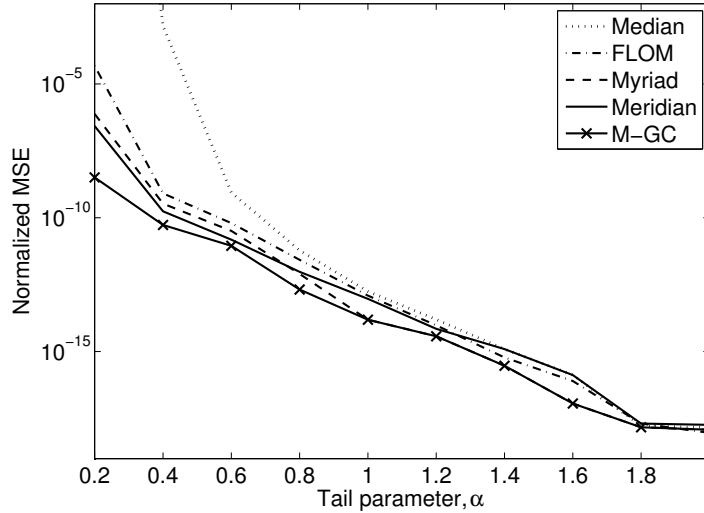


Figure 3.5: Power line communication enhancement. MSE for different filtering structures as function of the tail parameter α .

are plotted, as a function of α , in Fig. 3.5. The results show that the various methods perform somewhat similarly in the less demanding light-tailed noise environments, but that the more robust methods, in particular the M-CG approach, significantly outperforms in the heavy-tailed, impulsive environments. The time-domain results are presented in Fig. 3.6, which clearly show that the M-GC is more robust than the other operators, yielding a cleaner signal with fewer outliers and well preserved signal (symbol) transitions. The M-GC filter benefits from the optimization of the scale and tail parameters, and therefore perform at least as good as the myriad and meridian filters. Similarly, the M-GC filter performs better than the FLOM filter, which is widely used for processing stable processes [154].

3.5.2 Robust Blind Decentralized Estimation

Consider next a set of K distributed sensors, each making observations of a deterministic source signal θ . The observations are quantized with one bit (binary observations) and then these binary observations are transmitted through a noisy

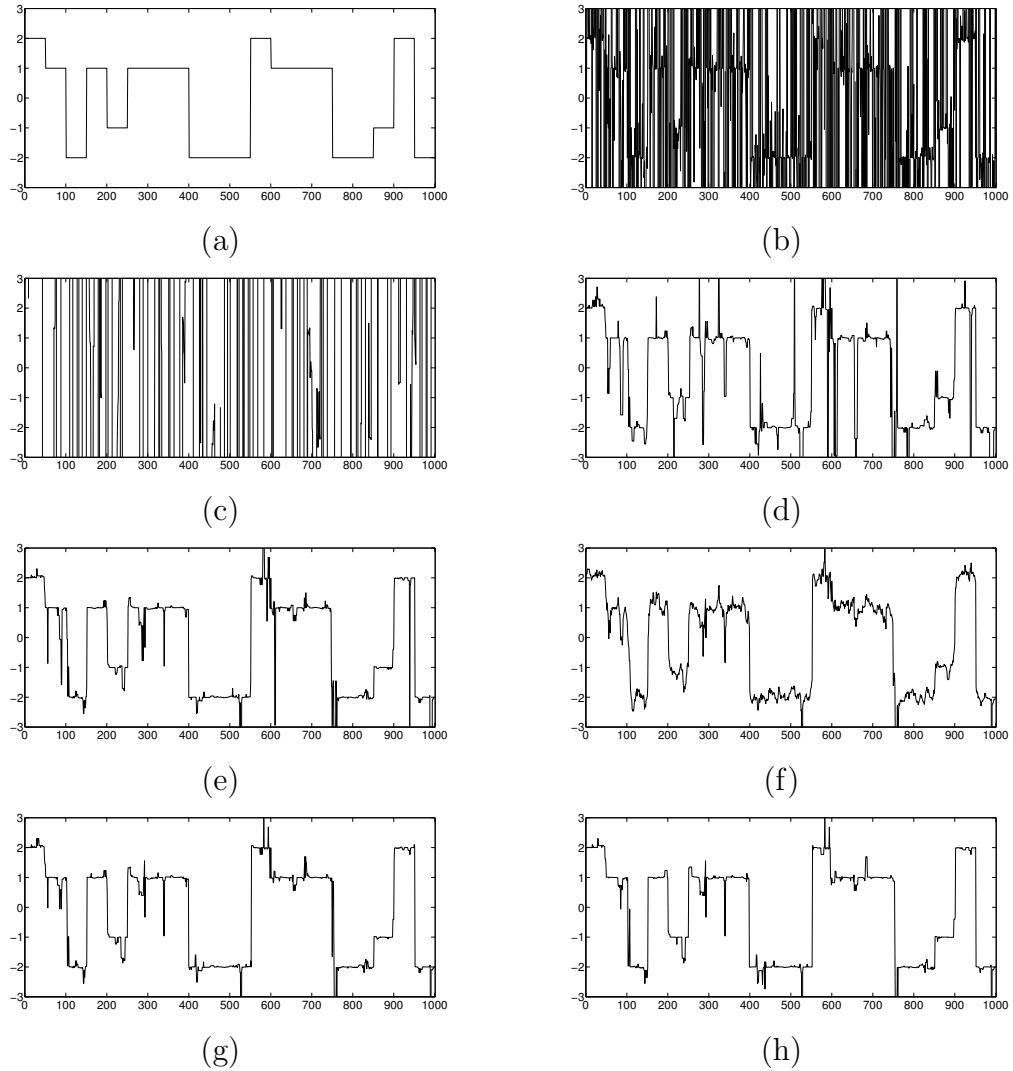


Figure 3.6: Power line communication enhancement. (a) Transmitted signal, (b) Received signal corrupted by α -stable noise $\alpha = 0.4$ Filtering results with: (c) Mean, (d) Median, (e) FLOM $p = 0.25$ (f) Myriad, (g) Meridian, (h) M-GC.

channel to a fusion center where θ is estimated (see [11, 12] and references therein). The observations are modeled as $x = \theta + n$, where n are sensor noise samples assumed to be zero-mean, spatially uncorrelated, independent and identically distributed. Thus the quantized binary observations are

$$b_k = \mathbf{1}\{x_k \in (\tau, +\infty)\} \quad (3.45)$$

for $k = 1, 2, \dots, K$, where τ is a real valued constant and $\mathbf{1}\{\cdot\}$ is the indicator function. The observations received at the fusion center are modeled by

$$y = (2b - 1) + w = m + w \quad (3.46)$$

where w are zero-mean independent channel noise samples and the transformation $m_k = 2b_k - 1$ is made to adopt a binary phase shift keying (BPSK) scheme.

The channel noise density function is denoted by $w_k \sim f_w(u)$. When this noise is impulsive (*e.g.*, atmospheric noise or underwater acoustic noise), traditional Gaussian-based methods (*e.g.*, least squares) do not perform well. We extend the blind decentralized estimation method proposed in [11], modeling the channel corruption as GCD noise and deriving a robust estimation method for impulsive channel noise scenarios. The sensor noise, n , is modeled as zero-mean additive white Gaussian noise with variance σ_n^2 , while the channel noise, w , is modeled as zero-location additive white GCD noise with scale parameter σ_w and tail constant p . A realistic approach to the estimation problem in sensor networks assumes that the noise pdf is known but that the values of some parameters are unknown [11]. In the following, we consider the estimation problem when the sensor noise parameter σ_n is known and the channel noise tail constant p and scale parameter σ_w are unknown.

Instrumental to the scheme presented is the fact that b_k is a Bernoulli random

variable with parameter

$$\psi(\theta) \triangleq \Pr\{b_k = +1\} = 1 - F_n(\tau - \theta) \quad (3.47)$$

where $F_n(\cdot)$ is the cumulative distribution function of n_k . The PDF of the noisy observations received at the fusion center is given by

$$f_y(y) = \psi(\theta)f_w(y - 1) + [1 - \psi(\theta)]f_w(y + 1). \quad (3.48)$$

Note that the resulting pdf is a GCD mixture with mixing parameters ψ and $[1 - \psi]$. To simplify the problem, we first estimate $\psi = \psi(\theta)$ and then utilize the invariance of the ML estimate to determine θ using (3.47).

Using the log-likelihood function, the ML estimate of $\psi \in (0, 1)$ reduces to

$$\hat{\psi} = \arg \max_{\psi} \sum_{k=1}^K \log\{\psi f_w(y_k - 1) + [1 - \psi]f_w(y_k + 1)\}. \quad (3.49)$$

The unknown parameter set for the estimation problem is $\{\psi, \sigma_w, p\}$. We address this problem utilizing the well known EM algorithm [129] and a variation of Algorithm 3 in section 3.3.4. The followings are the *E*- and *M*- steps for the considered sensor network application.

E-step: Let the parameters estimated at the j -th iteration be marked by a superscript (j) and $\Gamma^{(j)} = (\hat{\sigma}_w^{(j)}, \hat{p}^{(j)})$. The posterior probabilities are computed as

$$q_k = \frac{\hat{\psi}^{(j)} f_w(y_k - 1 | \Gamma^{(j)})}{\hat{\psi}^{(j)} f_w(y_k - 1 | \Gamma^{(j)}) + [1 - \hat{\psi}^{(j)}] f_w(y_k + 1 | \Gamma^{(j)})}. \quad (3.50)$$

M-step: The ML estimates $\{\hat{\psi}^{(j+1)}, \Gamma^{(j+1)}\}$ are given by

$$\hat{\psi}^{(j+1)} = \frac{1}{K} \sum_{k=1}^K q_k, \quad \text{and,} \quad \Gamma^{(j+1)} = \arg \max_{\Gamma} \Lambda(\Gamma) \quad (3.51)$$

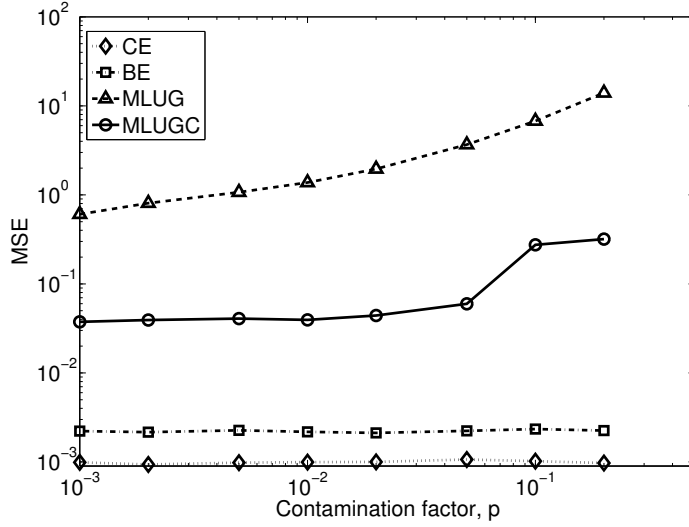
where

$$\Lambda(\Gamma) = \sum_{k=1}^K q_k \Upsilon(y_k - 1; \Gamma) + (1 - q_k) \Upsilon(y_k + 1; \Gamma) \quad (3.52)$$

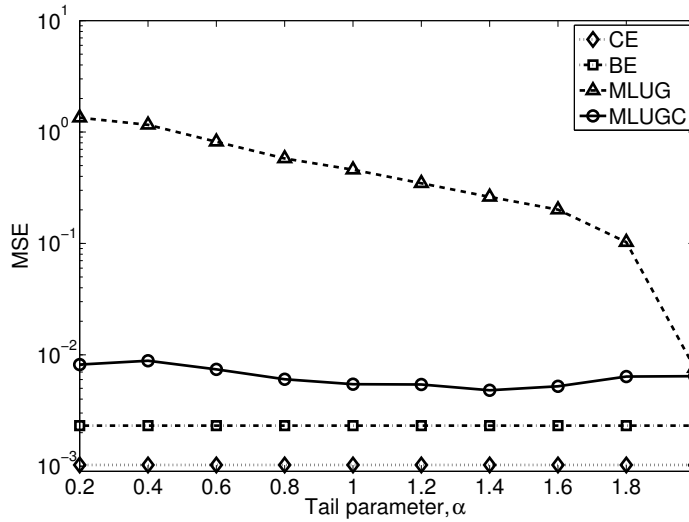
where $\Upsilon(u; \Gamma) = \log a(p) + \log \sigma_w - 2p^{-1} \log(\sigma_w^p + |u|^p)$ and $a(p) = p\Gamma(2/p)/2(\Gamma(1/p))^2$. We use a suboptimal estimate of p in this case, choosing the value from $P = \{0.5, 1, 1.5, 2\}$ that maximizes (3.51).

Numerical results comparing the derived GCD method, coined maximum likelihood with unknown generalized Cauchy channel parameters (MLUGC), with the Gaussian channel based method derived in [11], referred to as maximum likelihood with unknown Gaussian channel parameter (MLUG), are presented in Fig. 3.7. The MSE is used as a comparison metric. As a reference, the MSE of the binary estimator (BE) and the clairvoyant estimator (CE) (estimators in perfect transmission) are also included.

A sensor network with the following parameters is used: $\theta = 1$, $\tau = 0$, $\sigma_n = 1$ and $K = 1000$ and the results are averaged for 200 independent realizations. For the channel noise we use two models: contaminated p -Gaussian and α -stable distributions. Fig. 3.7 (a) shows results for contaminated p -Gaussian noise with the variance set as $\sigma_w^2 = 0.5$ and varying p (percentage of contamination) from 10^{-3} to 0.2. The results show a gain of at least an order of magnitude over the Gaussian-derived method. Results for α -stable distributed noise are shown in Fig. 3.7 (b), with scale parameter $\sigma_w = 0.5$ and the tail parameter, α , varying from 0.2 to 2 (very impulsive to Gaussian noise). It can be observed that the GCD-derived method has a gain of at least an order of magnitude for all α . Furthermore, the MLUGC method has a nearly constant MSE for the entire range. It is of note that the MSE of the MLUGC method is comparable to that obtained by the MLUG (Gaussian-derived) for the especial case when $\alpha = 2$ (Gaussian case), meaning that the GCD-derived method is robust under heavy-tailed and light-tailed environments.



(a)



(b)

Figure 3.7: Sensor network example with parameters: $\theta = 1$, $\tau = 0$, $\sigma_n = 1$ and $K = 1000$. Comparison of MLUGC, MLUG, BE and CE. (a) Channel noise contaminated p -Gaussian distributed with $\sigma_w^2 = 0.5$. MSE as function of the of the contamination parameter, p . (b) Channel noise α -stable distributed with $\sigma_w = 0.5$. MSE as function of the tail parameter, α .

3.5.3 Robust Clustering

As a final example, we present a robust fuzzy clustering procedure based on the LL_p metrics defined in Section 3.4, which is suitable for clustering data points involving heavy-tailed non-Gaussian processes. Dave proposed the *noise clustering* (NC) algorithm to address noisy data in [74, 75]. The NC approach is successful in improving the robustness of a variety of prototype-based clustering methods. This method considers the noise as a separate class and represents it by a prototype that has a constant distance δ .

Let $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^N$, $\mathbf{x}_j \in \mathbb{R}^n$, be a finite data set and C the given number of clusters. NC partitions the data set by minimizing the following function, proposed in [74]:

$$J(\mathbf{Z}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m d(\mathbf{x}_j, \mathbf{z}_i) + \sum_{j=1}^N \delta (1 - \sum_{i=1}^C u_{ij})^m \quad (3.53)$$

where $\mathbf{Z} = [\mathbf{z}_1; \dots; \mathbf{z}_C]$ is a matrix whose rows are the cluster centers, $m \in (1, \infty)$ is a weighting exponent, and, $d(\mathbf{x}_j, \mathbf{z}_i)$ is the squared L_2 distance from a data point \mathbf{x}_j to the center \mathbf{z}_i . $\mathbf{U} = [u_{ij}]$ is a $C \times N$ matrix, called a constraint fuzzy partition of \mathcal{X} , which satisfies [74]

$$u_{ij} \in [0, 1] \quad \forall i, j, \quad 0 < \sum_{j=1}^N u_{ij} < N \quad \forall i \quad \text{and} \quad \sum_{i=1}^C u_{ij} < 1 \quad \forall j. \quad (3.54)$$

The u_{ij} weight represents the membership of the i -th sample to the j -th cluster. Minimization of the objective function with respect to \mathbf{U} , subject to the constraints in (3.54), gives [74]

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left[\frac{d(\mathbf{x}_j, \mathbf{z}_i)}{d(\mathbf{x}_j, \mathbf{z}_k)} \right]^{1/(m-1)} + \left[\frac{d(\mathbf{x}_j, \mathbf{z}_i)}{\delta} \right]^{1/(m-1)}} \quad (3.55)$$

Compared with the basic fuzzy C-means (FCM), the membership constraint is relaxed to $\sum_{i=1}^C u_{ij} < 1$. The second term in the denominator of (3.55) becomes large for outliers, thus yielding small membership values and improving robustness of prototype-based clustering algorithms.

To further improve robustness, we propose the application of LL_p metrics in the NC approach. Substituting the LL_p norm for d in (3.53) yields the objective function

$$J(\mathbf{Z}) = \sum_{i=1}^C \sum_{j=1}^N (u_{ij})^m \|\mathbf{x}_j - \mathbf{z}_i\|_{LL_{p,\sigma}} + \sum_{j=1}^N \delta(1 - \sum_{i=1}^C u_{ij})^m \quad (3.56)$$

Given the objective function $J(\mathbf{Z})$, a set of vectors $\{\mathbf{z}\}_{i=1}^N$ that minimize $J(\mathbf{Z})$ must be determined. As in FCM, fix-point iterations are utilized to obtain the solution. We use a variation of the fixed point recursion proposed in Section 3.3.4 to achieve this goal. Differentiating $J(\mathbf{Z})$ with respect to each dimension l of \mathbf{z}_s , treating the u_{ij} terms as constants, and setting it to zero yields the fixed point function. Thus the recursion algorithm can be written as

$$z_{sl}(t+1) = \frac{\sum_{j=1}^N w_j(t) x_{jl}}{\sum_{j=1}^N w_j(t)} \quad (3.57)$$

with

$$w_j(t) = \frac{u_{sj}^m p |x_{jl} - z_{sl}(t)|^{p-2}}{\sigma^p + |x_{jl} - z_{sl}(t)|^p} \quad (3.58)$$

where t denotes the iteration number. The recursion is terminated when $\|z_s(t+1) - z_s(t)\|_2 < \epsilon$ for some given $\epsilon > 0$. This method is used to find the update of the cluster centers. Alternation of (3.55) and (3.57) gives an algorithm to find the cluster centers that converge to a local minimum of the cost function.

In the NC approach, $m = 1$ corresponds to crisp memberships, and increasing m represent increased fuzziness and soft rejection of outliers. When m is too large,

spurious cluster may exist. The choice of the constant distance δ also influences the fuzzy membership; if it is too small, then we can not distinguish good clusters from outliers, and if it is too large, the result diverges from the basic FCM. Based on [74], we set $\delta = (\lambda/N^2) \sum_{i \neq j}^N \|\mathbf{x}_i - \mathbf{x}_j\|_{LL_p, \sigma}$, where λ is a scale parameter. In order to reduce the local minimum caused by initialization of the NC approach, we use classical k -means on a small subset of the data to initialize a set of cluster centers. The proposed algorithm is summarized in Algorithm 4 and is coined the LL_p based Noise Clustering (LL_p -NC) algorithm.

Algorithm 4 LL_p based Noise Clustering Algorithm

Require: cluster number C , weighting parameter m , δ , maximum number of iterations or terminate parameter ϵ .

- 1: Initialize cluster centers.
- 2: **while** $\|z_s(t+1) - z_s(t)\|_2 > \epsilon$ or a maximum number of iterations is not reached **do**
- 3: Compute the fuzzy set \mathbf{U} using (3.55) and
- 4: Update cluster centers using (3.57).
- 5: **end while**
- 6: **return** Cluster centroids $\mathbf{Z} = [\mathbf{z}_1; \dots; \mathbf{z}_C]$.

Experimental results show that for multi-group heavy-tailed process, the results of the LL_p based method generally converges to the global minimum. However, to address the problem of local minima, the clustering algorithm is performed multiple times with different random initializations (subsets randomly sampled) and with a fixed small number of iterations. The best result is selected as the final solution.

Simulations to validate the performance of GCD based clustering algorithm (LL_p -NC) in heavy tailed environments are carried out and results summarized in Table 3.2. The experiment uses three synthetic data sets of 400 points each with different distributions and 100 points in each cluster. The locations of the centers for the three sets are: $[-6, 2]$, $[-2, -2]$, $[2, 4]$ and $[3, 0]$ for each set. The first set has Cauchy distributed clusters (GCD, $p = 2$) with $\sigma = 1$, and is shown in Fig. 3.8. The second

Table 3.2: Clustering results for GCD processes and α -stable process

N	MSE	MAD	LL_p	Average Distance
LL_p -NC	0.34987	0.62897	0.0968	Cauchy
L_1 -NC	1.8186	1.8361	0.1262	15.39
Similarity-based	1.6513	1.136	0.18236	
LL_p -NC	0.85197	0.9283	0.1521	Meridian
L_1 -NC	5.887	2.7311	0.5573	50.363
Similarity-based	5.2309	2.4627	1.8416	
LL_p -NC	0.50408	0.73618	0.1896	α -stable
L_1 -NC	3.2105	2.7684	0.2174	44.435
Similarity-based	1.7578	1.6322	1.0112	

has the meridian distribution (GCD, $p = 1$), with $\sigma = 1$. The meridian is a very impulsive distribution. The third set has a two dimensional α -stable distribution with $\alpha = 0.9$ and $\gamma = 1$, which is also a very impulsive case. The algorithm was run 200 times for each set with different initializations, setting the maximum number of iterations to 50, $\epsilon = 0.0001$, and $\lambda = 0.1$.

To evaluate the results, we calculate the MSE, the mean absolute deviation (MAD), and the LL_p distance between the solutions and the true cluster centers, averaging the results for 200 trials. The LL_p NC approach is compared with classical NC employing the L_1 distance and the similarity-based method in [169]. The average L_2 distance between all points in the set (AD) is shown as a reference for each sample set. As the results show, GCD based clustering outperforms both traditional NC and similarity-based methods in heavy-tailed environments. Of note is the meridian case, which is a very impulsive distribution. The GCD clustering results are significantly more accurate than those obtained by the other approaches.

3.6 Concluding Remarks

This Chapter presents a GCD based theoretical approach that allows the formulation of challenging problems in a robust fashion. Within this framework, we

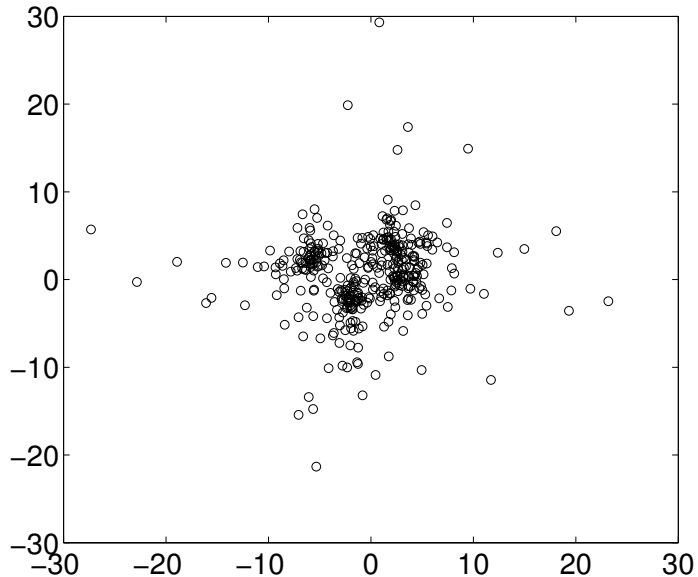


Figure 3.8: Data set for clustering example 1: Cauchy distributed samples with cluster centers $[-6,2]$, $[-2,-2]$, $[2,4]$ and $[3,0]$.

establish a statistical relationship between the GGD and GCD families. The proposed framework, due to its flexibility, subsumes GGD based developments, thereby guaranteeing performance improvements over the traditional problem formulation techniques. Properties of the derived techniques are analyzed. Three particular applications are developed under this framework: 1) robust filtering for power line communications, 2) robust estimation in sensor networks with noisy channels and 3) robust fuzzy clustering. Results from the applications show that the proposed GCD-derived methods provide a robust framework in impulsive heavy-tailed environments, with performance comparable to existing methods in less demanding light-tailed environments.

Chapter 4

ROBUST SAMPLING AND RECONSTRUCTION METHODS FOR SPARSE SIGNALS IN THE PRESENCE OF IMPULSIVE NOISE

4.1 Introduction

Compressed sensing (CS) is a recently introduced novel framework that goes against the traditional data acquisition paradigm. CS demonstrates that a sparse, or compressible, signal can be acquired using a low rate acquisition process that projects the signal onto a small set of vectors incoherent with the sparsity basis [18, 37, 41, 79]. The fundamental CS premise is that certain classes of signals, such as natural images, have a concise representation in terms of a sparsity inducing basis where most of the coefficients are zero or small, and only few are significant. A signal is sampled taking a few linear measurements and subsequently recovered using an optimization formulation that determines the sparsest representation consistent with the measurements. One of the most attractive features of CS is that random vectors and randomly selected vectors from orthonormal matrices are incoherent with any sparsity-inducing basis, with high probability [19, 37, 40, 79, 130], therefore allowing easy construction of sensing matrices.

Since noise is always present in real data acquisition systems, a range of different algorithms and methods have been developed that enable approximate reconstruction of sparse signals from noisy compressive measurements [2, 35, 42, 44, 60, 65,

83,101,115,134,135,139,149,158,161]. The reconstruction quality for such compressible signals is proportional to that of the signal's optimal sparse approximation and the magnitude of the noise. Noise aware algorithms follow three basic approaches: geometric-based algorithms [42, 44, 65, 83, 158], greedy algorithms [134, 135, 161], or complexity based algorithms [2, 60, 115]. Most such algorithms provide bounds for the L_2 reconstruction error based on the assumption that the corrupting noise is bounded, Gaussian, or, at a minimum, has finite variance.

Noise contributions to the overall system can be separated into two models: *observation* noise and *sampling* noise [148]. Consider first the case of observation noise. Observation noise is any perturbation introduced to the underlying signal *prior* to the sampling process, *e.g.*, channel noise effects in communications or salt and pepper noise in images. The (additive) model of the signal in this case is:

$$x = x_0 + w, \tag{4.1}$$

where $x_0 \in \mathbb{R}^n$ is the original signal and w is the additive noise.

Sampling noise, in contrast, introduces perturbations to the measurements in conjunction with the sampling process, *i.e.*,

$$y = y_0(x) + z \tag{4.2}$$

where $y_0(x) \in \mathbb{R}^m$, $m < n$, is the vector of samples, or measurements of x , and z is the corrupting noise, *e.g.*, quantization noise or sensor noise. If we consider linear measurements as in the traditional CS literature, then, in the noiseless case, $y = \Phi x_0$, where Φ is the measurement matrix and, for a noisy signal,

$$y = \Phi x = \Phi x_0 + r \tag{4.3}$$

with $r = \Phi w$. When w is Gaussian, r is also Gaussian yielding (4.1) and (4.2) similar; therefore, they can be approached by the same methods and use linear measurements and Gaussian-derived reconstruction algorithms. However, when the signal is corrupted by gross errors or heavy-tailed impulsive noise, linear measurements are severely degraded, with original signal information masked by large amplitude samples spread throughout the measurements as a result of the linear sampling process. This introduction of large valued corrupting samples, and their spreading across measurements, causes traditional reconstruction algorithms to fail in their attempts to recover a fair approximation of the underlying signal.

Another tenet of traditional reconstruction algorithms that fails in demanding environments is the assumption that the sampling noise has finite variance. If a corrupting process has infinite, or even very large, variance, the allowable, and likely resulting, reconstructions will be far from the desired original signal. Recent works have begun to address the reconstruction of sparse signals from measurements corrupted by outliers, *e.g.* due to missing data in the measurement process; or in the context of channel coding, due to transmission problems [39, 43, 145]. Popilka *et. al* proposed a reconstruction algorithm based on the sparsity of the measurement error pattern to estimate first the error, and then estimate the true signal, in an iterative process [145]. A similar approach is followed by Candès *et. al*, but in the context of error correction coding, where the number of measurements (codeword length) is larger than the dimension of the signal (original sequence) [39, 43] and the codeword is assumed to be corrupted by gross outliers. A drawback of this approach is that the reconstruction relies on the error sparsity to first estimate the error, but if the sparsity condition is not met, the performance of the algorithm degrades, or many iterations may be required to yield a fair estimate.

Notably, there exists a broad spectrum of applications where practice has shown non-Gaussian, heavy-tailed processes emerge. Examples of such applications

are: wireless communications, teletraffic, hydrology, geology, atmospheric noise, economics and image and video processing (see [4,24] and the references therein). Thus, the motivation is clear for developing robust CS techniques that address these challenging environments.

The contributions of this Chapter are the development of: 1) robust information operators, $I_m : \mathbb{R}^n \rightarrow \mathbb{R}^m$, that sample m pieces of information from x in the presence of observation noise and 2) robust reconstruction algorithms, $A_n : \mathbb{R}^m \rightarrow \mathbb{R}^n$, that render approximate reconstructions of original sparse signals from small sets of measurements, when (possibly) heavy-tailed noise is introduced in the sampling process. It is well known that nonlinear methods, derived from heavy-tailed distributions, overcome the limitations of traditional linear signal processing methods in the presence of such signals [4, 24]. We approach the problem of impulsive observation and sampling noise from a statistical point of view and propose methods based on robust statistics [118], specifically methods derived from the Generalized Cauchy distribution (GCD) family [7, 9, 48, 106–108]. For the case of impulsive observation noise we propose a more robust nonlinear measurement operator, based on the weighed myriad estimators family [4]. The myriad measurement offers robustness in impulsive environments, thereby decreasing the effect of impulsive noise while, at the same time, allowing the use of standard reconstruction algorithms derived for linear measurements. To recover sparse signals from impulsive noise introduced in the measurement process, we propose a geometric approach based on robust estimation theory. The proposed non-convex program seeks a solution that minimizes the L_1 norm subject to a nonlinear constraint based on the Lorentzian norm, thereby defining a feasible set that diminish the effect of gross errors and consequently performing a denoising effect.

The organization of the Chapter is as follows: In Section 4.2 we present a brief review of CS and sparse reconstruction methods noting their limitations in impulsive

environments. In Section 4.3 we present the problem in which the observation noise is impulsive. The so called myriad measurements are defined and their properties are discussed along with the approach's capabilities as a measurement method for CS. In Section 4.4 a robust reconstruction algorithm is proposed and its performance is analyzed. Numerical results for the proposed methods are presented for a variety of impulsive models in Section 4.5. Finally, we conclude in Section 4.6 with closing thoughts and future directions.

4.2 Background and Motivation

This section gives a brief review of the CS problem and, geometric and greedy reconstruction approaches. Next we present an analysis of current Least Squares (LS) based methods noting their limitations in the presence of impulsive noise. Explanatory examples are presented also for both impulsive observation and sampling noise.

4.2.1 Compressed Sensing Review

Let $x \in \mathbb{R}^n$ be a signal that is either s -sparse or compressible in some orthogonal basis Ψ . The signal is s -sparse if only s of its coefficients are nonzero, where $s \ll n$. The signal is compressible if its ordered set of coefficients decay rapidly and x can be well-approximated by just the first s coefficients. Thus $x = \Psi^T \theta$, where $\theta \in \mathbb{R}^n$ is the vector of coefficients.

Let $\{\phi_i\}_{i=1}^m$ be a set of measurements vectors that are incoherent with the sparsity basis. Incoherence indicates that none of the vectors $\{\phi_i\}_{i=1}^m$ have a sparse or compressible representation in the original sparsity basis Ψ [40]. The signal x is measured by taking projections on to the set $\{\phi_i\}_{i=1}^m$. The measurement process is a linear map $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$, where $m < n$, and $y = \Phi x$ is the vector containing all the measurements. If we set $\Xi = \Phi \Psi^T$, then the measurement vector becomes $y = \Xi \theta$. For example, the vectors $\{\phi_i\}_{i=1}^m$ can be random vectors with independent entries

or vectors randomly chosen from an orthogonal basis. In the following we assume, without loss of generality, that $\Psi = \mathbf{I}$, the canonical basis for \mathbb{R}^n , yielding $x = \theta$.

The ideal recovery of x from the measurements y is achieved by the following problem

$$\min_{x \in \mathbb{R}^n} \|x\|_0 \quad \text{subject to} \quad \Phi x = y, \quad (4.4)$$

which finds the sparsest vector x consistent with the measurements. The problem in (4.4) is combinatorial and almost surely intractable; however, it can be relaxed into a convex problem if the measurement matrix Φ satisfies certain conditions [43], which are described below. The convex relaxation is

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad \Phi x = y, \quad (4.5)$$

which can be solved by linear programming techniques. The optimization problem in (4.5) is known as basis pursuit and was previously used to find sparse representations on over complete dictionaries [65]. We focus on the results of [43], which shows that if x is s -sparse, and Φ obeys a restricted isometry property (RIP), then the solution of (1) is also the solution of (4.4). Letting Φ be a sensing matrix with normalized columns, in the L_2 sense, and T be a subset of indices of $\{1, \dots, n\}$, the definition of the restricted isometry constants is as follows.

Definition 8. For every integer $1 \leq q \leq n$ let define δ_q as the q -restricted isometry constant of Φ as the smallest positive quantity such that

$$(1 - \delta_q)\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1 + \delta_q)\|v\|_2^2 \quad (4.6)$$

for all subsets T of cardinality at most q and vectors v supported on T .

If $\delta_q \in [0, 1)$, a RIP requires that every set of columns with cardinality less than q approximately behaves like an orthonormal system. It is shown in [38] that

if $\delta_{2s} < \sqrt{2} - 1$ the solution of (4.5) recovers any sparse signal with support size of at most s . It is also shown that random matrices with Gaussian or sub-Gaussian entries have restricted isometry constants in the interval $[0, 1)$ with high probability provided that $m = O(s \log(n))$ [19].

In a more realistic scenario the measurements are corrupted with noise and can be modeled as $y = \Phi x + r$, where r is additive zero-mean white noise. In the presence of noise, variations of the aforementioned strategies have been shown to reliably approximate the signal, assuming that certain *a priori* information is known about the signal or the noise process. The results not only apply for strictly s -sparse signals but also to compressible signals.

Basis Pursuit with L_2 constraint relaxes the requirement that the reconstructed signal explain exactly the measurements [38, 42]. The reconstruction solves the optimization problem

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad \|y - \Phi x\|_2 \leq \epsilon, \quad (4.7)$$

for some small $\epsilon > 0$. In [38] it is shown that if $\|r\|_2 \leq \epsilon$ and if $\delta_{2s} < \sqrt{2} - 1$, then the reconstructed signal \hat{x} is guaranteed to obey

$$\|x - \hat{x}\|_2 \leq C\epsilon, \quad (4.8)$$

where the constant C depends on δ_{2s} . The Lasso [158] and Basis Pursuit Denoising [65] are two alternative formulations of the problem in (4.7). The Dantzig Selector is a similar convex program for statistical estimation proposed in [44] that uses an L_∞ constraint instead of L_2 , reconstruction error also depends on the noise variance.

Other approaches used to find a sparse solution employ greedy algorithms that iteratively construct a sparse approximation to the signal. Such algorithms include Matching Pursuit (MP) [128], Orthogonal Matching Pursuit (OMP) [161]

and their derivations [134, 135]. Matching Pursuit is a greedy algorithm that iteratively incorporates in the reconstructed signal the component from the measurement set that explains the largest portion of the residual from the previous iteration. Orthogonal Matching Pursuit additionally orthogonalizes the residual against all measurement vectors selected in previous iterations. The number of measurements required for OMP is also $O(s \log(n))$ for Gaussian measurement matrices [161]. The algorithm stops when the residual reaches a magnitude below a set threshold. The conditions for proper termination involve knowledge of the signal sparsity or the noise variance to achieve the desired denoising effect [35].

4.2.2 Impulsive Noise in CS

Recall that the noise contributions can be separated into two models: observation noise and sampling noise. In the following, we make an analysis of traditional sampling operators based on linear projections and traditional reconstruction algorithms based on LS methods in impulsive environments. We use the oracle estimator to derive the best performance that can be achieved with LS derived methods when no prior information about the distribution of the original signal is known. See for example [42, 44] for similar analysis. The oracle estimator is called the ideal estimator because the support of x_0 (the set of positions of the s non zero coefficients of x_0 , $\Omega \subset \{1, \dots, n\}$) is known in advance. Using this prior information and assuming Gaussian distributed errors, we can construct an estimator by using the least squares projection on to the subspace spanned by the columns of Φ with indices in Ω .

Lets consider first the case when we have observation noise $x = x_0 + w$ and take w to be a vector of i.i.d. random variables. Define each sample as

$$y_i = \sum_{j=1}^n \phi_{ij} x_j, \quad i = 1, \dots, m.$$

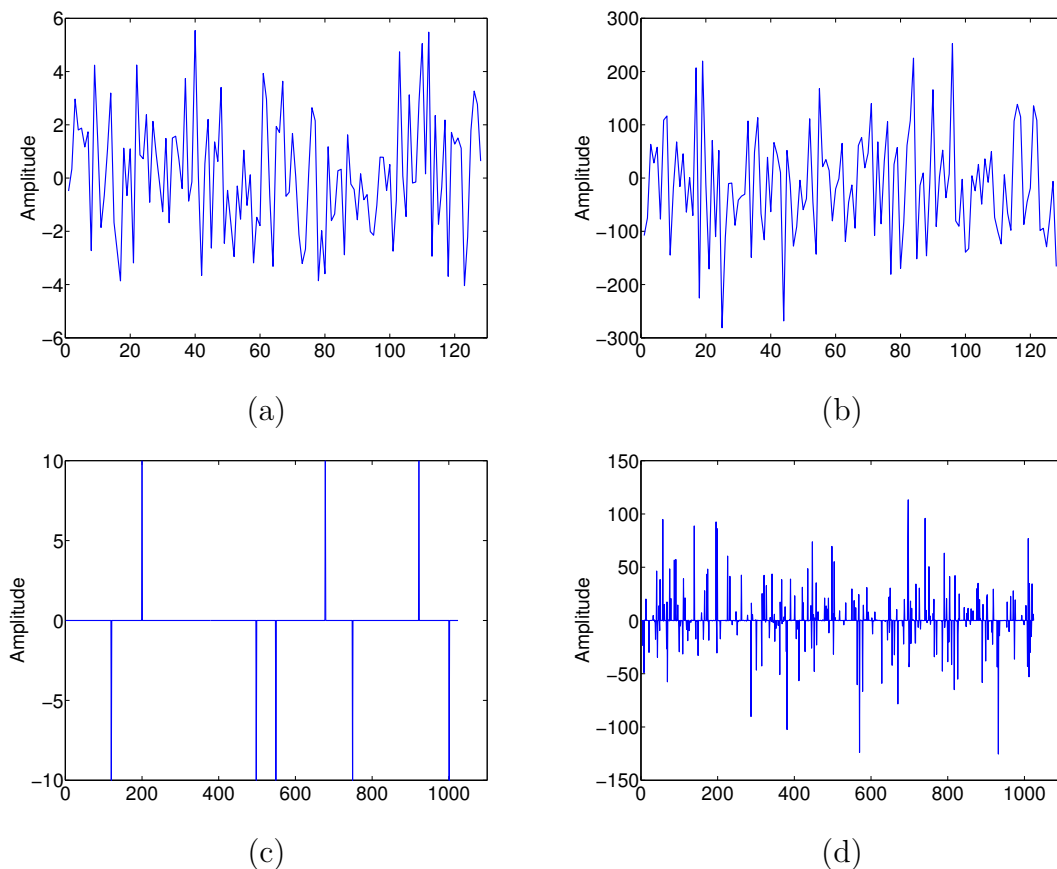


Figure 4.1: Example of a signal corrupted by a single outlier. (a) Linear projections in the noiseless case. (b) Linear projections when the signal is corrupted with a single impulse. (c) Original sparse signal. (d) Reconstructed sparse signal from linear projections using BP with L_2 constraint.

Then the sampling operator becomes $y = \Phi x_0 + z$ where $z = \Phi w$. If w is a Gaussian process then z is also Gaussian and all the methods described in 4.2.1 recover a fair approximation of x_0 provided $\mathbb{E}(w_j^2)$ is small, and the transformation Φ is stable so $\mathbb{E}(z_i^2)$ is small also. If the noise w is not Gaussian and, furthermore it is an impulsive process, linear measurements are severely affected because the large amplitude of the noise components spread throughout every measurement. In the presence of gross errors all the reconstruction algorithms mentioned above fail because the variance of all z_i is very large or not finite. A common example of this phenomena in image

processing is salt and pepper noise. The mean square error (MSE) of the oracle estimator in this case is

$$\mathbb{E}\|x^* - x_0\|_2^2 = \mathbb{E}(w_1^2) \|(\Phi_\Omega^T \Phi_\Omega)^{-1} \Phi_\Omega^T \Phi\|_F^2 \quad (4.9)$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix and $\mathbb{E}(w_1^2)$ is the common second moment for all w_i 's. Since the support of the signal is known by the oracle estimator, its MSE is the lowest reconstruction error that can be achieved by all methods described above (LS based without prior knowledge about the signal). Given the finite variance constraint, we can see that linear projections are not the best sampling operators to use when the underlying signal is corrupted by impulsive noise. Consider the example in Fig. 4.1, which employs a signal sparse in a Hadamard basis of dimension $n = 1024$. The sparsity level is $s = 8$ and the signal is measured through 256 linear projections with a Gaussian matrix. In Fig. 4.1 (a) we show the linear projections y in the noiseless case; the reconstruction from these samples is shown in Fig. 4.1 (c). Now we add a single outlier to the original signal of amplitude $\delta = 10^3$. The position of the outlier is randomly chosen. In Fig. 4.1 (b) the linear projections for the signal corrupted with the impulse are shown; the reconstructed signal from these projections is shown in Fig. 4.1 (d). Here BP and BP with L_2 constraint were used as the reconstruction algorithms for the noiseless case and the corrupted case respectively. The reconstruction SNR for the noiseless case is 229.5 dB and for the corrupted case is -25.7 dB. As can be seen in Fig. 4.1 (b) the large amplitude of the outlier spreads through all the samples thus making almost impossible for BP with L_2 constraint to recover the original signal.

Consider now the sampling noise case, $y = \Phi x + z$. Suppose we have an oracle estimator, then a lower bound for the MSE is given by

$$\mathbb{E}\|x^* - x_0\|_2^2 = \mathbb{E}(z_1^2) \text{Tr}\{(\Phi_\Omega^T \Phi_\Omega)^{-1}\} \geq \frac{s\mathbb{E}(z_1^2)}{1 + \delta_s} \quad (4.10)$$

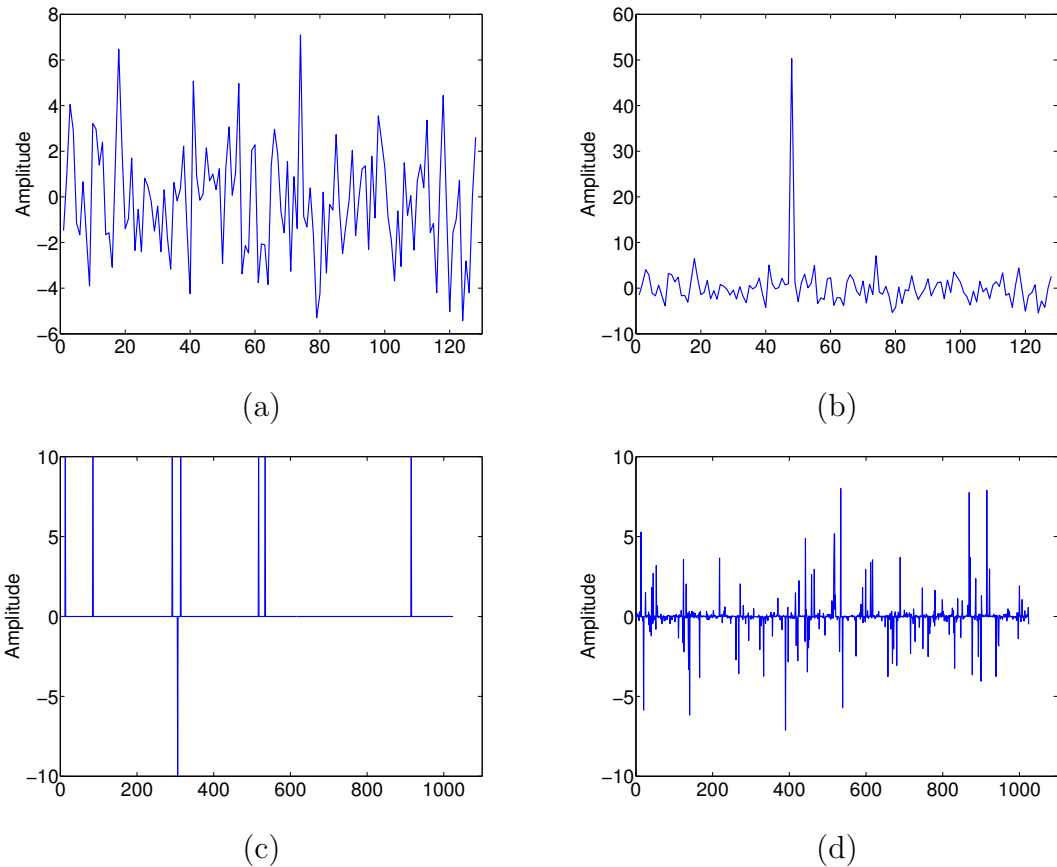


Figure 4.2: Example of measurements corrupted by a single outlier. (a) Linear projections in the noiseless case. (b) Linear projections corrupted with one impulse. (c) Original sparse signal. (d) Reconstructed sparse signal using BP with L_2 constraint.

where δ_s is the restricted isometry constant of Φ_Ω and $\mathbb{E}(z_1^2)$ is the common second moment for all z_i 's. This estimator's reconstruction error depends on $\mathbb{E}(z_1^2)$ since LS regression is derived from Gaussian assumptions. When the noise is Gaussian or otherwise has bounded variance, the expected error is also finite and the oracle estimator can yield an approximate reconstruction. Traditional CS reconstruction algorithms reviewed in 4.2.1 (without prior information about the signal) are based on LS methods (Gaussian noise assumption) and thus have the oracle estimator as a theoretical bound and, importantly, depend on the finite variance assumption. In

the case of impulsive heavy-tailed noise corrupted measurements, the variance may be very large or even infinite, thereby leading to a large reconstruction error even for this ideal estimator. Fig. 4.2 shows an example of a sparse signal sensed by 128 linear projections with a Gaussian measurement matrix. The measurements are corrupted by a single outlier of amplitude 50. In Fig. 4.2 (a) we show the uncorrupted samples and in Fig. 4.2 (b) the corrupted samples. The reconstruction from the uncorrupted samples is shown in Fig. 4.2 (c) and Fig. 4.2 (d) shows the reconstruction from the corrupted samples. As in the last example, BP was used for the noiseless case and BP with L_2 constraint was used for the corrupted case. The reconstruction SNR for the noiseless case is 193.1 dB and -5.7 dB for the corrupted case.

Since LS based methods do not achieve good performance in impulsive environments, we make use of robust statistics to find more appropriate methods to address the problem of impulsive noise in CS. Specifically, we utilize methods derived from the algebraic-tailed Generalized Cauchy distribution (GCD) family developed in Chapter 3.

4.3 Robust Sampling Functions

Of interest here is the design of an information operator $I_m : \mathbb{R}^n \rightarrow \mathbb{R}^m$ that samples m pieces of information of x in a fashion that: (a) allows faithful reconstruction and (b) is immune to outlier corruption. Consider a signal $x_0 \in \mathbb{R}^n$ that is sparse in some basis Ψ (for the sake of simplicity we set $\Psi = I$) and the signal model

$$x = x_0 + w,$$

where x is the noisy observed signal and w is white noise. Defining each sample as

$$y_i = f(\phi_i, x), \tag{4.11}$$

where ϕ_i are the sampling kernels (rows of the sensing matrix Φ), the information operator takes the form

$$y = I_m(x) = (f(\phi_1, x), f(\phi_2, x), \dots, f(\phi_m, x)). \quad (4.12)$$

Suppose $A_n : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a reconstruction strategy that recovers x_0 from y . The goal is then to design information operators such that traditional CS LS-based reconstruction strategies (using y) can achieve a close reconstruction. In the presence of impulsive noise, the information operator, $I(x)$, should have the following desired properties:

- P1. $I_m(x)$ should remove the influence of gross errors in order to preserve the basic information of x_0 .
- P2. $I_m(x) \rightarrow \Phi x_0$ asymptotically ($w \rightarrow 0$), *i.e.*, the measurements should be as similar as possible to the linear measurements in the noiseless case.

P1 states the necessity of I_m to sample true information of x_0 and not information of the noise. P2 characterized the desired output of the sampling function so that y contains the essential information of x_0 , as in the noiseless case, so that current CS theory and reconstruction methods can be effectively applied to these samples. This property is a practical constraint, applied so as to allow use of traditional CS reconstruction algorithms. If other reconstruction strategies are allowed, then linear sampling functions are not necessarily the desired output.

From these two properties, it follows that the desired operator I_m should behave like a robust estimator of the correlation between the sampling kernel and the original signal.

4.3.1 Myriad Projections

Problems associated with linear sampling in the presence of outliers are noted above. Based on these problems and the desired properties of a robust sampling operator, we propose the weighted myriad estimator as a robust nonlinear operator f to sample the information of x . There are multiple reasons to use the weighted myriad as a sampling operator, including its robustness to impulsive noise and asymptotic linearity property, which guarantees the sampling capabilities in the noiseless case. These properties allow us to tune the linearity parameter and use standard reconstruction algorithms, thereby treating the reconstruction as a light-tailed noisy problem. As the weighted myriad estimator (especial $p = 2$ case of the M-GC estimator defined in Chapter 3) is utilized in the subsequent development, we employ the common notation for this as:

$$\hat{\theta} = \text{myriad}(\sigma, |h_i| \circ \text{sgn}(h_i)x_i) \Big|_{i=1}^n \quad (4.13)$$

where \circ denotes the weighting operation for the myriad estimator [4].

We begin with the definition of the myriad projections.

Definition 9. *Let $\Phi \in \mathbb{R}^{m \times n}$ be a given measurement matrix and ϕ_{ij} its ij -th entry. Then the myriad projections are defined as*

$$f_K(\phi_i, x) = a_i \cdot \text{myriad}(K; |\phi_{ij}| \circ \text{sgn}(\phi_{ij})x_j) \Big|_{j=1}^n \quad (4.14)$$

for $i = 1, \dots, m$, where $a_i = \sum_{j=1}^n |\phi_{ij}|$ is a scaling factor introduced to preserve the amplitude of the measurement.

We now state the basic properties of myriad measurements and relate them to P1 and P2. A desirable property in robust measurements is that outlier samples have minimal influence on the projections (P1). The following property states the outlier rejection property of the myriad measurements.

Property 5. *Let $K < \infty$, then*

$$\lim_{x_n \rightarrow \pm\infty} f_K(\phi_i, x_1, x_2, \dots, x_n) = f_K(\phi_i, x_1, x_2, \dots, x_{n-1}). \quad (4.15)$$

The following property states the asymptotic behavior of weighted myriad measurements and is based on the linearity property of the weighted myriad estimator.

Property 6. *In the limit as $K \rightarrow \infty$, the weighted myriad measurement reduces to a linear projection on to ϕ_i . This is*

$$\lim_{K \rightarrow \infty} f_K(\phi_i, x) = \sum_{j=1}^n \phi_{ij} x_j. \quad (4.16)$$

Thus in the limiting case, as $K \rightarrow \infty$, myriad measurements meet property P2 and can be used as robust sampling functions or robust correlation measures.

Note that the properties above follow from myriad operator properties [108]. Also, the weighted myriad measurement converges to a selection type estimator as $K \rightarrow 0$ [4] and, in the limiting case when $K = 0$, the measurement becomes independent of the weight vector (rows of Φ), converging to the most repeated value in the set. Thus, in this limiting case, the recovery process can not return the true signal because of the loss of information.

4.3.2 Asymptotical analysis and parameter tuning

Since the myriad operator is the ML estimator of location for standard Cauchy samples, the myriad location estimator is Gaussian distributed [106]. A remark is that this property is not automatically inherited by weighted myriad filters for all signals models, but it provides a model for the myriad measurements.

Letting $\eta \sim \mathcal{N}(0, \nu)$, we can model the myriad measurements as

$$f_K(\phi_i, x) = \sum_{j=1}^n \phi_{ij} x_j + r_i, \quad i = 1, \dots, m, \quad (4.17)$$

where $r_i \rightarrow \eta$ in distribution, as $n \rightarrow \infty$. The variance, ν , is the asymptotic variance, and depends on the strength of the corrupting process. The following proposition gives the asymptotic variance of the myriad estimator for the standard Cauchy case.

Proposition 3. *Let X be a standard Cauchy random variable with location parameter θ and scale parameter σ , then the asymptotic variance of the myriad estimator is given by $\nu = 2\sigma^2$.*

Proof. The asymptotic variance for M-estimators is given by $\mathbb{E}(\psi^2)/[\mathbb{E}(\psi')]^2$, where ψ is the influence function of the estimator [118]. For the myriad estimator $\psi(x) = 2x/(\sigma^2 + x^2)$, then taking the expectation with respect to the standard Cauchy distribution gives

$$\mathbb{E}[\psi^2(X)] = \frac{\sigma}{\pi} \int_{-\infty}^{\infty} \frac{4x^2}{(\sigma^2 + x^2)^3} dx = \frac{1}{2\sigma^2}$$

and

$$\mathbb{E}[\psi'(X)] = \frac{\sigma}{\pi} \int_{-\infty}^{\infty} \frac{2\sigma^2 - 2x^2}{(\sigma^2 + x^2)^3} dx = \frac{1}{2\sigma^2},$$

leading to the desired result. See [106] for further details. \square

Notice that ν is finite and smaller than the variance of the original algebraic-tailed noise [107], since the second moment is not defined for Cauchy random variables. This result allows us to use current LS-based reconstruction algorithms, designed for linear projections, with myriad projection inputs.

The availability of the tuning parameter K provides myriad projections with a variety of modes of operations that range from highly impulse resistant measurements to linear projections. However, there exists a tradeoff between linearity and

robustness of the sampling operator, which is controlled by K . Large K values lead to good approximations to linear measurements, but yield results less resistant to outliers. Small K values make the myriad measurement robust to impulsive noise, but the measurements are highly nonlinear, leading to degradations in reconstruction.

Determining the optimal K (optimal in the sense that the measurements are as close as possible to the noiseless linear case) from the corrupted signal is still an open question. In [4] it is observed that setting K as the sample range, $x_{(1)} - x_{(0)}$ (where $x_{(q)}$ denotes the q -th quantile of x), often makes the myriad a fair approximation to a linear combination. On the other hand, setting K as half the interquartile range, $(x_{(0.75)} - x_{(0.25)})/2$, considers implicitly half the samples unreliable, giving resilience to gross errors. Therefore choosing a value of K between the sample range and half the interquartile range yields a value that is well behaved in both Gaussian and impulsive models. Experimental results show that a linearity parameter set as

$$K = \frac{x_{(0.875)} - x_{(0.125)}}{2}, \quad (4.18)$$

leads to good performance in both Gaussian and impulsive environments. Setting K in this range implicitly assume a signal with 25% of samples corrupted by outliers and 75% well behaved. This is demonstrated experimentally in Section 4.5.

An observation of note is that when the signal is sparse in the canonical basis and sparse-like impulsive noise is added directly, the signal and noise become undistinguishable, unless the noise has significantly larger amplitude. Fortunately, noise is generally introduced in the observation domain, which is rarely coincident with the sparsity inducing basis. Another observation to make is that myriad projections are more expensive in terms of computational resources since an optimization problem must be solved for each projection, whereas linear projections can be computed with lower cost or can even be observed directly. Thus, myriad projections

should be used to measure a signal when the sensing conditions are not ideal, *e.g.*, in noisy signal environment resulting from, for instance, overshoots in front end hardware (ADC before the CS measurement), or when robust sensing procedures are needed. One final observation is that if structured sampling matrices are utilized (*e.g.* sparse matrices that meet RIP [17]), the cost of computing myriad projections can be significantly lowered.

4.4 Robust Reconstruction Algorithms

This section addresses the problem of signal reconstruction from corrupted measurements. Let $x_0 \in \mathbb{R}^n$ be an s -sparse signal and $\Phi \in \mathbb{R}^{m \times n}$ a measurement matrix. Consider the measurement model

$$y = \Phi x_0 + z$$

where z is white additive noise. In this case the objective is to design robust reconstruction algorithms, $A_n : \mathbb{R}^m \rightarrow \mathbb{R}^n$, that yield approximate reconstructions of the original sparse signal from a small set of measurements, assuming linear sampling operators, in the presence of (probably impulsive) sampling noise.

The reconstruction strategies need to be robust and stable in the sense that small variations in the noiseless samples should yield small variations in the reconstructed signal, even when a fraction of the samples are corrupted by gross errors. Most of current reconstruction algorithms use the L_2 norm as the metric for the residual error; but as detailed in Section 4.2, the L_2 norm is not an appropriate metric when the samples are corrupted by outliers. Using this arguments, we propose to use a robust metric to penalize the residual and address the impulsive sampling noise problem.

4.4.1 Lorentzian constrained L_1 minimization

Using the strong theoretical guarantees of L_1 minimization for sparse recovery of underdetermined systems of equations (see [38, 80] for example), we propose the following non-linear constrained optimization problem to estimate a sparse signal from the noisy measurements y :

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad \|y - \Phi x\|_{LL_2, \gamma} \leq \epsilon, \quad (4.19)$$

where $\|u\|_{LL_2, \gamma}$ is the Lorentzian norm (LL_p norm with $p = 2$). The L_1 objective encourages sparsity in the solution (as in other geometric approaches [42, 44, 65]) and the Lorentzian constraint controls the residual error. Thus the intuition behind utilizing a Lorentzian norm defined feasible set is the construction of a search space that is not severely affected by sparse large outliers, but which also behaves as an L_2 ball for small Gaussian-like errors. Further justifying use of the Lorentzian norm is the existence of logarithmic moments for algebraic-tailed distributions, as second moments are infinite or not defined for such distributions and therefore not an appropriate measure of process strength [109].

The main result of this section is given by Theorem 6 below. Lets assume that all z_i , $i = 1, \dots, m$, are i.i.d random variables with common distribution $f_Z(z)$ and $Z \sim f_Z(z)$. The result shows that the solution to (6.1) is a sparse signal with an L_2 error that is dependent on the logarithmic moment $\mathbb{E} \log\{1 + (Z/\gamma)^2\}$. Note that the dependence on the noise logarithmic moment, rather than its second order moment, makes the formulation in (6.1) robust and stable to algebraic-tailed and impulsive corrupted samples, *i.e.* we exploit the fact that $\mathbb{E}\|z\|_{LL_2, \gamma} < \infty$ while $\mathbb{E}\|z\|_2$ might not be.

Theorem 6. *Let Φ be a sensing matrix such that $\delta_{2s} < \sqrt{2} - 1$. Then for any signal x_0 such that $|T_0| \leq s$, where $T_0 = \text{supp}(x_0)$, and observation noise z with*

$\|z\|_{LL_2,\gamma} \leq \epsilon$, the solution to (6.1), x^* , obeys the following bound

$$\|x^* - x_0\|_2 \leq C_s \cdot 2\gamma \cdot \sqrt{m(e^\epsilon - 1)}, \quad (4.20)$$

where the constant C_s depends only on δ_{2s} .

Proof. Lets decompose x^* as $x^* = x_0 + h$. We are going to divide the proof in two parts: first find an upper bound for $\|\Phi h\|_2$ and second show that $\|h\|_2 \approx \|\Phi h\|_2$ up to a constant.

Define u, v as $u = \Phi x^* - y$, and $v = y - \Phi x_0$. Since x^* is a feasible point and the error is assumed to obey $\|z\|_{LL_2,\gamma} \leq \epsilon$, it follows that $\|u\|_{LL_2,\gamma} \leq \epsilon$ and $\|v\|_{LL_2,\gamma} \leq \epsilon$. Then

$$\begin{aligned} \|\Phi h\|_2 &\stackrel{(a)}{\leq} \|u\|_2 + \|v\|_2 \\ &\stackrel{(b)}{\leq} \gamma \sqrt{m(e^{\|u\|_{LL_2,\gamma}} - 1)} + \gamma \sqrt{m(e^{\|v\|_{LL_2,\gamma}} - 1)} \\ &\stackrel{(c)}{\leq} 2\gamma \sqrt{m(e^\epsilon - 1)}, \end{aligned} \quad (4.21)$$

where (a) follows from the triangle inequality, (b) from Lemma 2 in Chapter 3 with $p = 2$, and (c) from the Lorentzian bounds on u and v .

It just remains to show that $\|h\|_2 \approx \|\Phi h\|_2$. It was shown in [38], in the proof of theorem 1.2, that if $\delta_{2s} < \sqrt{2} - 1$ then

$$\|h\|_2 \leq \frac{\sqrt{2}\|\Phi h\|_2 \sqrt{1 + \delta_{2s}}}{(1 - \delta_{2s} - \sqrt{2}\delta_{2s})}. \quad (4.22)$$

Finally replacing (4.21) in to (4.22) we have

$$\|h\|_2 \leq \frac{\sqrt{2 + 2\delta_{2s}}}{(1 - \delta_{2s} - \sqrt{2}\delta_{2s})} 2\gamma \sqrt{m(e^\epsilon - 1)}.$$

which is the desired result. The condition $\delta_{2s} < \sqrt{2} - 1$ is a necessary condition for

the constant C_s to be positive. □

The constant C_s is given by $C_s = \sqrt{2 + 2\delta_{2s}}(1 - \delta_{2s} - \sqrt{2\delta_{2s}})^{-1}$ and is rather small for reasonable values of δ_{2s} . One remark on (4.20) is that as $\epsilon \rightarrow 0$ the reconstruction error goes to zero and in the noiseless case ($\epsilon = 0$) the reconstruction is perfect. The \sqrt{m} factor in (4.20) represents the dependence of the reconstruction error on the noise vector size (norm), since this one scales with m . This dependence is implicit in the error bound of BP with L_2 constraint, equation (4.8) in section 4.2.1, since ϵ depends on m (see [42]).

Notably, γ controls the robustness of the employed norm and ϵ the radius of the feasibility set LL_2 ball. Details on the estimation of these parameters and an analysis in the standard Cauchy model are given below.

4.4.2 Analysis under the Cauchy model

To facilitate the reconstruction quality analysis, we consider the ideal case when the sampling noise is standard Cauchy distributed. Suppose an oracle estimator is available, which knows the support of the original signal in advance. Define Ω as the support of the original signal and denote as $x_\Omega \in \mathbb{R}^s$ the restriction of x to Ω . Let z be a vector of i.i.d. Cauchy random variables with location parameter $\theta = 0$ and dispersion parameter σ . Lets denote by Z a random variable with the same distribution as the noise. The ML estimate of x_Ω in this case is:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^s} \|y - \Phi_\Omega \beta\|_{LL_2, \sigma}. \quad (4.23)$$

The estimate derived in (4.23) is a robust regressor that is optimal for the standard Cauchy model and the generalized t -student distribution [16, 106]. Moreover, the approach has proven to be effective in general impulsive environments, as well as light-tailed (Gaussian) environments [106].

It is known from robust statistics that asymptotic theory of M-estimators can be extended to robust regressors [118]. Let $\rho(x)$ denote the cost function of the estimator and $\psi(x) = \rho'(x)$ its influence function. In the case of ML estimates, $\rho(x) = -\log f(x)$, where $f(x)$ is the probability density of the samples. It can be proven that, asymptotically (as $s/m \rightarrow 0$):

$$\mathbb{E}\|\hat{\beta} - x_\Omega\|_2^2 = \frac{\mathbb{E}(\psi^2)}{[\mathbb{E}(\psi')]^2} \text{Tr}\{(\Phi_\Omega^T \Phi_\Omega)^{-1}\}, \quad (4.24)$$

if the matrix Φ_Ω is of rank s ; $\rho(x)$ is continuous and nonmonotone; $\psi(x)$ is continuous, bounded and $\mathbb{E}(\psi(Z)) = 0$ [118]. The term $\mathbb{E}(\psi^2)/[\mathbb{E}(\psi')]^2$ is the asymptotic variance of the M-estimator. It can be easily verified that the cost function and influence function for the standard Cauchy ML estimator meet each condition mentioned above (see for example Appendix C of [106]) and remember that for the standard Cauchy case, the asymptotic variance of the myriad estimator is $2\sigma^2$. Notice that if Φ satisfies the RIP of order s then Φ_Ω will approximately behave as an orthonormal system and, for the Cauchy case (4.24) can be lower-bounded by

$$\mathbb{E}\|\hat{\beta} - x_\Omega\|_2^2 \geq \frac{2s\sigma^2}{1 + \delta_s} \quad (4.25)$$

where the inequality comes because the eigenvalues of $(\Phi_\Omega^T \Phi_\Omega)$ lie in the interval $[1 - \delta_s, 1 + \delta_s]$. This lower bound provides an asymptotical result of the best performance that can be achieved using Lorentzian-based regressors under the standard Cauchy model.

With the Cauchy model as a reference we can derive estimates for the proper value of γ and ϵ to maximize performance. Again assume that z is a vector of Cauchy random variables with location parameter $\theta = 0$ and scale parameter σ . We make use of the following result for standard Cauchy random variables.

Lemma 3. *Let X be a standard Cauchy random variable with location parameter $\theta = 0$ and scale parameter σ , then:*

$$\mathbb{E} \log\{1 + \gamma^{-2}X^2\} = 2 \log\left(1 + \frac{\sigma}{\gamma}\right). \quad (4.26)$$

Proof. Recall that

$$\int_{-\infty}^{\infty} \frac{\log\{a^2 + p^2x^2\}}{1 + x^2} dx = 2\pi \log(a + p), \quad a, p > 0.$$

Then

$$\begin{aligned} \mathbb{E} \log\{1 + \gamma^{-2}X^2\} &= \frac{1}{\pi\sigma} \int_{-\infty}^{\infty} \frac{\log\{1 + \gamma^{-2}x^2\}}{1 + \sigma^{-2}x^2} dx \\ &= 2 \log\left(1 + \frac{\sigma}{\gamma}\right). \end{aligned}$$

□

Using Lemma 3 we can see that $\mathbb{E}\|z\|_{LL,\gamma} = m\mathbb{E} \log\{1 + \gamma^{-2}z_i^2\} = 2m \log(1 + \gamma^{-1}\sigma)$. If we use this expected value as an upper bound for the level of noise we can tolerate, then $\epsilon = 2m \log(1 + \gamma^{-1}\sigma)$ and, replacing this value in (4.20), the upper bound for the reconstruction error becomes

$$\|x^* - x_o\|_2 \leq C_s \cdot 2\gamma \cdot \sqrt{m} \left[\left(1 + \frac{\sigma}{\gamma}\right)^{2m} - 1 \right]^{\frac{1}{2}}. \quad (4.27)$$

From (4.27) it can be noticed that the reconstruction depends on the value of σ and γ . Here σ is the scale parameter of the Cauchy distribution and it is a measure of the strength of the noise, thus as $\sigma \rightarrow 0$ the error decreases. On the other hand, γ is a scale parameter for the Lorentzian norm and it controls the outlier resilience. A proper scale parameter is one that, makes the Lorentzian norm behave as an L_2 norm for errors smaller than the typical amplitude of the uncorrupted measurements;

therefore, we propose to use an estimate of scale of y_0 (uncorrupted samples) and set γ as the Median Absolute Deviation (MAD) of y . Thus γ is simply set as a robust estimate of scale, which makes the higher order polynomial terms in (4.27) vanish in the case $\gamma \gg \sigma$, and the error approximate $C_s \cdot 2m\sqrt{\sigma\gamma}$. Thus the ratio σ/γ can be interpreted as a noise to signal ratio (NSR), with the closer the value is to 0, the better expected reconstruction. The noise scale parameter, σ , is assumed to be *a priori* information known by the reconstruction algorithm.

4.4.3 Debiasing

Once an approximate solution, \hat{x} , is obtained using the minimization in (6.1), we perform a debiasing step. This step consist of performing a regression on a subset of indexes of \hat{x} using the robust regressor in (4.23). The subset is defined as $I = \{i : |\hat{x}_i| > \alpha\}$ for some threshold $\alpha > 0$. Let $\tilde{x}_I \in \mathbb{R}^d$ be defined as

$$\tilde{x}_I = \arg \min_{x \in \mathbb{R}^d} \|y - \Phi_I x\|_{LL_2, \xi} \quad (4.28)$$

where $d = |I|$. The final estimated signal after the regression, \tilde{x} , is defined as \tilde{x}_I for those indexes in the subset I and zero outside I . Experimental results show that setting α as $\lambda \max_i |\hat{x}_i|$, where $0 < \lambda < 1$, yields good results in the reconstruction. In our experiments we use $\lambda = 0.1$. The parameter ξ in the Lorentzian norm in (4.28) is a scale parameter for the noise distribution (not be confused with γ in (6.1)), and is assumed to be *a priori* information of the sampling noise.

In summary, the problem in (6.1) selects the support of x , while the debiasing step chooses the optimal values for these components, based on a minimum Lorentzian criterion. The reconstruction algorithm composed of solving (6.1) and followed by the debiasing step is referred to as Lorentzian BP in the remainder of the paper. It is worth to point out that debiasing is not always desirable, since shrinking the selected coefficients can mitigate unusually large noise deviations [78].

Thus, in the presence of highly impulsive noise, this desirable effect may be undone by debiasing.

4.5 Experimental Results

This section illustrates the effectiveness of myriad measurements and Lorentzian BP as robust techniques for CS by means of numerical experiments and their comparison with standard CS linear sampling functions and de-noising algorithms. For all the experiments we create synthetic sparse signals, setting the length of the signal to $n = 1024$ and the cardinality of the sparse support to $s = 8$. The nonzero coefficients are drawn from a Rademacher distribution and their position randomly chosen so that the average power of the signal is always fixed to 0.78. The number of random measurements is set to $m = 128$ unless otherwise specified or varied. The signals are measured using measurement matrices Φ that have i.i.d. entries drawn from a standard normal distribution with normalized columns. We average 200 repetitions of each experiment, with different realizations of the sparse supports, random measurement matrices, and additive noise. The reconstruction Signal to Noise Ratio (R-SNR) is used to measure performance. To test the robustness of the methods, we use two noise models: α -stable distributed noise and Gaussian noise plus gross sparse errors. The α -stable model is very popular for modeling processes with infinite variance because of the generalized central limit theorem, which states that the limiting distribution of a sum of i.i.d. random variables belongs to the α -stable class [4]. There are two α -stable cases of particular interest: for $\alpha = 1$ the distribution reduces to the standard Cauchy distribution, from which the proposed methods are derived; and $\alpha = 2$ yields the Gaussian distribution, from which *LS* methods are derived. The Gaussian noise plus gross sparse errors model is represented as $N = X + V$, where $X \sim \mathcal{N}(0, \sigma)$ and V is a discrete random variable with alphabet $v = \{0, \delta, -\delta\}$ and probabilities $\{1 - p, p/2, p/2\}$, respectively. We refer to this model as contaminated p -Gaussian noise for the remainder of the paper, as p

represents the amount of gross error contamination. This model is used to represent small perturbations with gross errors such as erasures of the desired signal [39,145], which mimics realistic scenarios in many signal processing applications.

4.5.1 Robust Sampling: Myriad Measurements

In the following we present experiments performed to validate the use of myriad projections as robust compressive measurements for signal recovery in the presence of impulsive environments using standard reconstruction algorithms. We use Basis Pursuit (BP) in the noiseless case, and Basis Pursuit Denoising (BPD) [42, 65] and Orthogonal Matching Pursuit (OMP) [161] in the noisy case. As *a priori* information, it is assumed that the noise tolerance is known for BPD (ϵ in equation (4.7) from section 4.2) and that the sparsity level is known for OMP. In the following a preceding M in the name of the reconstruction algorithm (*e.g.* M-BPD), indicates that the reconstruction is performed using myriad projections. To evaluate the performance, we first make examples to validate that myriad projections meet properties P1 and P2 in section 4.3. Next addressed is the problem of tuning K from the input signals, and validation of the proposed estimate. With an algorithmically set K , we proceed to evaluate the performance for different noise models and for different numbers of samples.

We start with an example of a single impulse added to the original signal to show the outlier rejection capability of myriad measurements (property P1 in section 4.3). The amplitude of the impulse is set to 10^3 and the reconstruction is performed using OMP for both linear and myriad projections. The R-SNR is -28.6 dB for the linear projections and 32.2 dB for myriad projections, using the K estimation method proposed in Section 4.3 and subsequently analyzed, which in this case yields $K = 1.25$. The results are depicted in Fig. 4.3.

Next we evaluate the validity of the hypothesis that myriad projections meet property P2 and compare myriad measurements with linear measurements in the

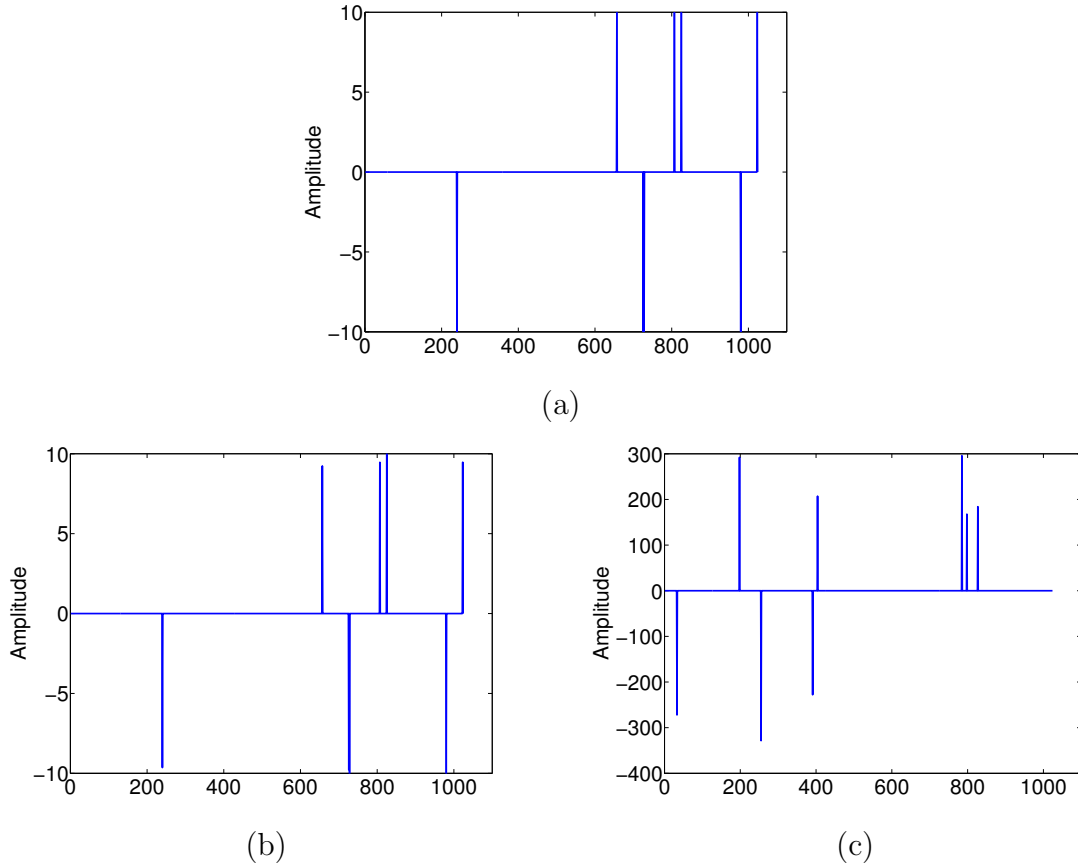


Figure 4.3: Outlier rejection example. (a) Original sparse signal. (b) Reconstructed signal from myriad projections, R-SNR=32.2 dB. (c) Reconstructed signal from linear projections, R-SNR=-28.6 dB.

noiseless case for several values of the linearity parameter K . The linearity parameter was varied over the range $[10^{-5}, 10^7]$ and the reconstruction SNR is used as comparison metric. Basis Pursuit (BP) and Orthogonal Matching Pursuit (OMP) are used as reconstruction algorithms for both myriad and linear measurements. Results are summarized in Fig. 4.4, which shows that myriad measurements yield fair signal reconstructions in the noiseless case as K increases and, more importantly, it shows that in the limit as $K \rightarrow \infty$ the performance of the reconstruction algorithms operating on myriad measurements approach the same performance of those supplied with linear measurements.

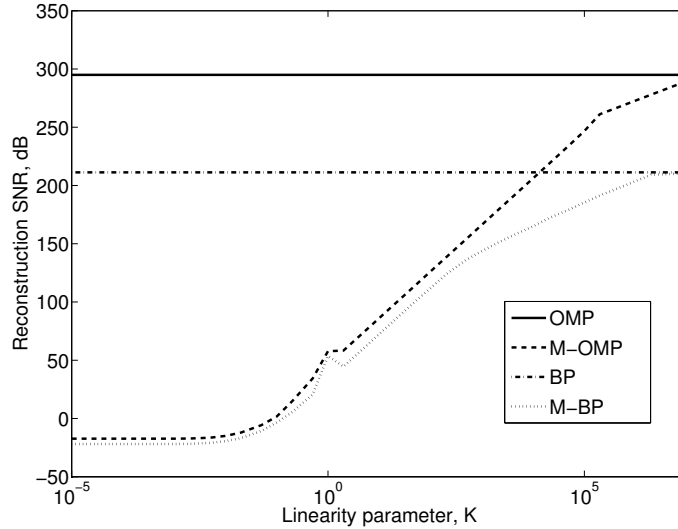


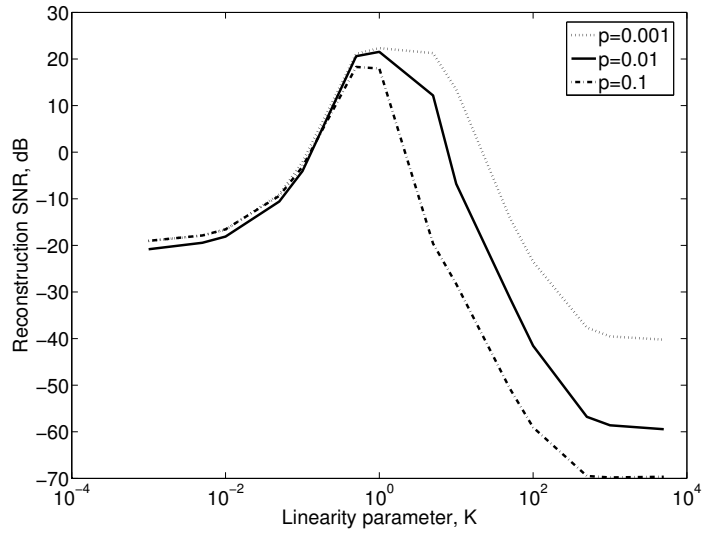
Figure 4.4: Comparison results between linear projections and myriad projections for the noiseless case, showing reconstruction SNR as a function of the linearity parameter, K . OMP and BP are used as reconstruction algorithms. The preceding M indicates that the reconstruction is performed using myriad projections.

Having established the applicability of myriad measurements in even noiseless cases, we now address the more demanding heavy-tailed environments. To explore the behavior of myriad measurements as a function of K , simulations using α -stable and contaminated p -Gaussian noise are performed. Three values for p are used, 0.001, 0.01, 0.1, with $\sigma^2 = 10^{-2}$ and $\delta = 10^3$ for the contaminated p -Gaussian model; and four values of α are used in the α -stable case, 0.5, 1, 1.5, 2, with scale parameter $\sigma = 0.1$. The results are summarized in Figs. 4.5 (a) and 4.5 (b), respectively. Consider the following observations. As $K \rightarrow 0$ the performance, in both cases, is degraded because the myriad filter tend to behave as a selection type estimator and, although it is robust to outliers, the reconstruction algorithms fail to render faithful reconstructions because of the non-linearities introduced into the projections. As K increases, the measurements become more linear and performance increases until maximum performance is achieved. From the maximum point, the

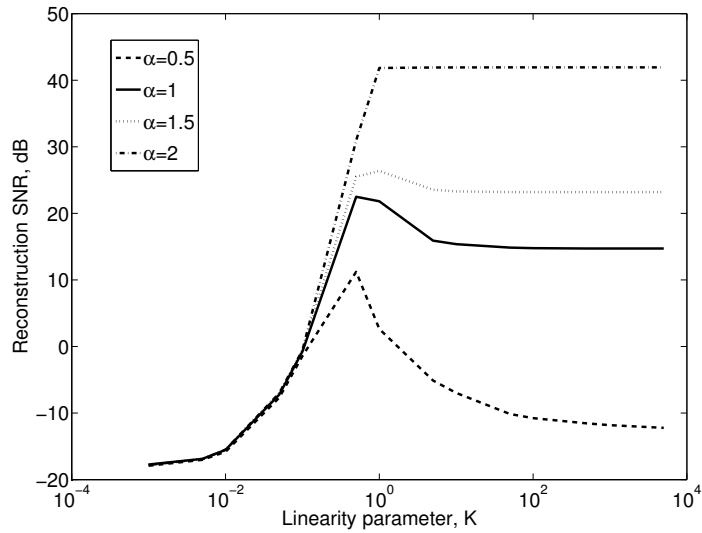
performance decreases as the measurements behave like linear measurements and exhibit diminishing robustness to outliers. In the case of contaminated p -Gaussian noise, the results are relatively invariant to p . The point of maximum performance is achieved in the same neighborhood of K for the three values of p , and similar R-SNR is achieved before that point. Beyond that point the performance is maintained for an interval depending on the impulsiveness of the contamination; $p = 0.001$ gives the longest interval and $p = 0.1$ the shortest. In the case of α -stable noise, the optimal K is largely independent of α , with $\alpha = 2$ representing the Gaussian special case in which performance is independent of K beyond the fixed maximum point.

As illustrated in the above experiment, the performance of myriad measurements as a sampling operator relies largely on the proper value of K . However, determining the optimal value of K , such that the measurements are as close as possible (in the L_2 sense) to the uncorrupted linear measurements, is still an open question. We propose to set K as in (4.18). Setting K to this value implicitly assumes a signal with 25% of samples corrupted by outliers and 75% well behaved. In the next experiment we make a comparison between the performance of myriad projections equipped with the optimal K and the signal-estimated K for standard Cauchy observation noise. The optimal K being found by exhaustive search. The normalized squared L_2 error between the uncorrupted linear measurements and myriad projections is used as a metric for comparison (normalized with respect to the L_2 norm of the uncorrupted measurements). The scale parameter of the Cauchy noise is varied from 10^{-2} to 10, giving a geometric signal to noise ratio¹(G-SNR) range of 44.4 dB through -15.4 dB respectively, to study the effect of noise strength. The

¹ The geometric SNR is defined as the ratio between the signal power and the noise geometric power, where the geometric power is a measure of strength for algebraic-tailed random variables whose second moments are not defined. The geometric power is defined as $S_0 = e^{\mathbb{E} \log |X|}$. See [109] for more details.



(a)



(b)

Figure 4.5: Reconstruction SNR as a function of the linearity parameter K for impulsive noise models. (a) Additive noise: contaminated p -Gaussian with p varying from 0.001 to 0.1. (b) Additive noise: α -S with α varying from 0.5 to 2.

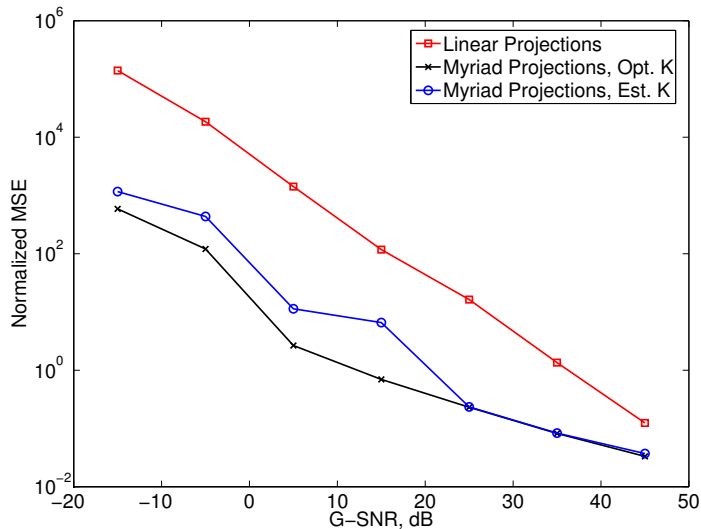


Figure 4.6: Myriad measurements performance comparison between optimal K and the proposed estimate for K . Normalized average MSE between myriad projections and clean linear projections for standard Cauchy noise. The scale parameter is varied from 10^{-2} to 10. The normalized MSE of corrupted linear projections is plotted for comparison.

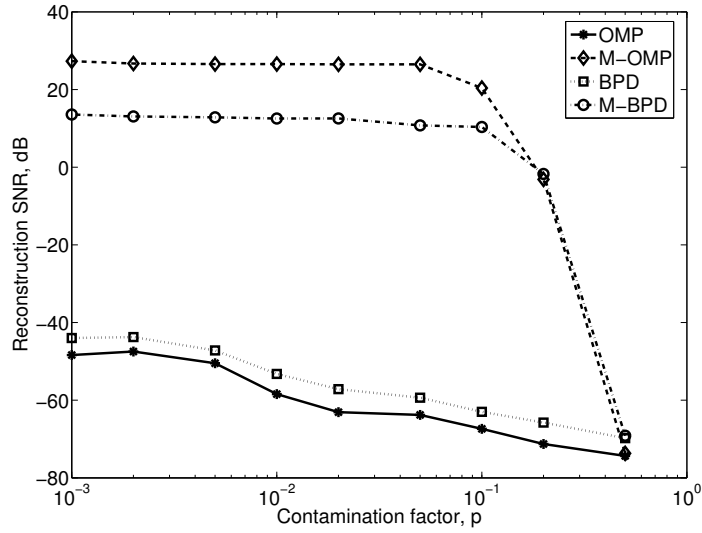
results are shown in Fig. 4.6. The normalized L_2 error for corrupted linear measurements is plotted as a reference. It can be noticed that for large G-SNR, *i.e.* small contaminations with 25 dB G-SNR and greater, the estimate of K achieves performance nearby relative to the optimum K . For G-SNR's below 25 dB, the myriad projection errors for estimated K becomes larger than the error achieved by the optimal K , but still are an order of magnitude smaller than linear projections in the worst case.

With a method for tuning the linearity parameter K from the corrupted signal we proceed to evaluate the performance of myriad projections in very impulsive environments. The next experiment shows how myriad projections compare to linear projections for two impulsive models: contaminated p -Gaussian and α -stable. For the contaminated p -Gaussian the variance of the Gaussian component is set as $\sigma^2 = 10^{-2}$, the amplitude of the gross errors as $\delta = 10^3$, and p was varied from

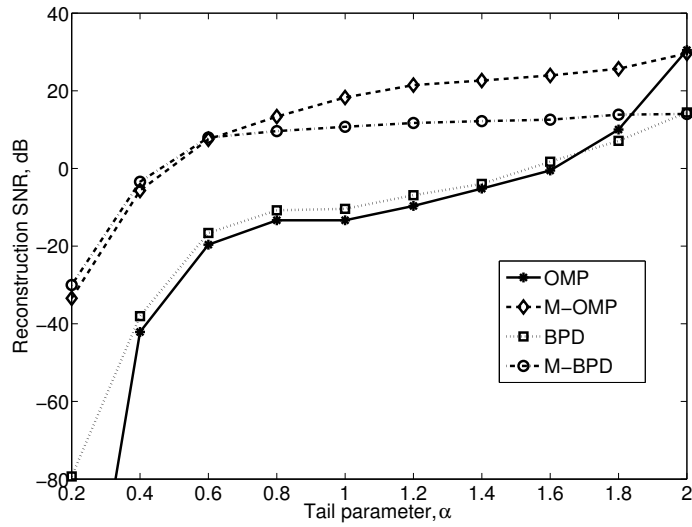
10^{-3} to 0.5. In the α -stable case, the scale parameter is set as $\sigma = 0.1$ and the tail parameter, α , is varied from 0.2 to 2. For BPD, the noise bound is set as $\epsilon = m\sigma^2$ for both noise models. Results for contaminated p -Gaussian noise are shown in in Fig. 4.7 (a) and results for α -stable noise are shown in Fig. 4.7 (b).

The results demonstrate that myriad projections-based reconstructions outperform linear projections-based reconstructions in the presence of heavy-tailed observation noise. Notably, in the case of contaminated p -Gaussian, myriad projections results are stable for a wide range of contamination factors, p , including contaminations of up to 10% of the signal's samples, making myriad projections a suitable sampling operator when samples are lost or erased. In the case of α -stable noise, both sampling operators perform poorly for small values of α but beyond $\alpha = 0.6$, myriad projections yield fair results with a R-SNR greater than 15 dB for both reconstruction algorithms tested. Of note is that in the Gaussian case ($\alpha = 2$), myriad projections based reconstruction is comparable with that of linear measurements based reconstruction.

As a practical experiment, we present an example utilizing a 256×256 image corrupted with salt and pepper noise with density of 0.01, *i.e.*, approximately 10% the pixels in the image are corrupted. We use a Haar basis as a sparsity inducing basis and a randomly sampled Hadamard matrix as the sampling matrix. The number of measurements, m , is set to $256 \times 256/4$ (25% of the number of pixels of the original image). As reconstruction algorithm we use BPD, the particular algorithm used is that described in [122]. The results are presented in Fig. 4.8, where (a) and (b) show the original image and the corrupted image respectively. Fig. 4.8 (c) shows the reconstructed image using linear projections as sampling functions and Fig. 4.8 (d) shows the reconstructed image using myriad projections. The reconstruction SNR is 11 dB and 23 dB for linear projections and myriad projections, respectively. Myriad projections remove the influence of the outliers (salt and pepper noise) in the



(a)



(b)

Figure 4.7: Comparison of linear projections with myriad projections for impulsive observation noise. (a) Contaminated p -Gaussian, R-SNR as a function of the contamination parameter, p . (b) α -S noise, R-SNR as a function of the tail parameter, α . OMP and BPD are used as reconstruction algorithms in both cases. The preceding M indicates that the reconstruction is performed using myriad projections.

input image, giving a gain of 12 dB in the reconstruction process. This example shows the utility of myriad projections when no prior information about the signal or corrupting noise is known.

As a final experiment for the noisy observation case, we evaluate the performance of myriad projections as the number of measurements varies from 16 (twice the sparsity level) to 512 (half the dimension of x), for a variety of impulsiveness levels. The results of linear projections based OMP are presented as a benchmark in Fig. 4.9 (a). We start with the noiseless case and then add α -stable noise with four different values of α , 2, 1.5, 1, 0.5, ranging from Gaussian noise to highly impulsive noise. The scale parameter of the noise is set as $\sigma = 0.1$ and the reconstruction algorithm used is OMP for all cases. The results are presented in Fig. 4.9 (b). In the noiseless case, myriad projections with a finite K cannot achieve the performance of linear projections (300 dB as shown in Fig. 4.4) due to the nonlinearity distortion introduced by the sampling process. However, in the Gaussian case myriad-based OMP achieves the same performance as linear-based OMP, *i.e.* requiring the same number of projections, thus showing that the performance is not affected by the nonlinearities in this case. The results also show that as the impulsiveness level increases (α decreases), the performance decreases, as expected, with OMP needing more samples to compensate for the introduced distortion. This is a fundamental tradeoff since linear sampling based methods also need more samples to address lower SNR scenarios. A conclusion to be drawn is that myriad projections offer robustness in heavy-tailed environments but offer the same performance in terms of number of samples required for reconstruction as compared to linear projections in light-tailed noisy cases.



(a)



(b)

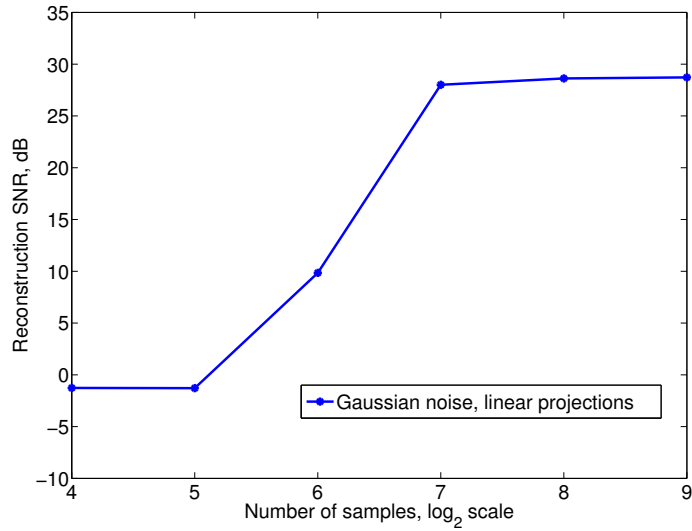


(c)

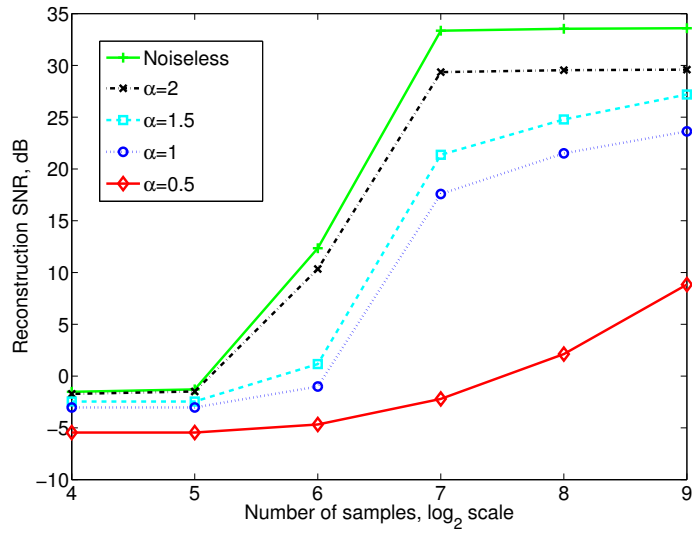


(d)

Figure 4.8: Example of a 256×256 image corrupted with salt and pepper noise with density 0.01. (a) Original image. (b) Noisy image. (c) Reconstructed image from linear projections using with BPD, R-SNR=11 dB. (d) Reconstructed image from myriad projections using BPD, R-SNR=23 dB.



(a)



(b)

Figure 4.9: Reconstruction SNR as a function of the number of measurements. (a) Linear projections based OMP with Gaussian observation noise. (b) Myriad-based OMP in the noiseless case and α -stable observation noise with α varying from 2 to 0.5.

4.5.2 Robust Reconstruction: Lorentzian BP

Next consider the case of corrupted measurements and performance evaluations of the Lorentzian BP reconstruction algorithm. The first experiment presented is a simple example with measurements corrupted by a single outlier. Then, we address the problem of estimating a proper value for the scale parameter of the Lorentzian norm, γ , from the corrupted measurements. With a proper γ , we test Lorentzian BP for different impulsive environments, starting with standard Cauchy sampling noise, from which Lorentzian BP is derived. As a final experiment, we test the performance of Lorentzian BP as a function of the number of samples for different noise environments. Basis Pursuit Denoising (BPD) and Orthogonal Matching Pursuit (OMP) were used as benchmarks. For both algorithms is assumed that the noise tolerance (ϵ) is known and OMP uses this tolerance as stop criteria.

A sequential quadratic programming (SQP) method is used to numerically solve the problem in (6.1). The method consists of three major steps at each iteration: Hessian approximation, solving a quadratic subproblem (QP), and performing a line search for the update. The approximation of the Hessian of the Lagrangian function is made using the BFGS updating method to have local information. The Hessian is then used to generate a quadratic subproblem, whose solution is used to form a search direction for a line search procedure. The line search method used is backtracking algorithm with a merit function. For further details see [136].

Consider first an example of measurements corrupted by a single outlier to show the outlier rejection capabilities of Lorentzian BP. The sparse signals employed in the previous subsection are again utilized, in this case with linear projections corrupted by a single 50 amplitude impulse. Also, BPD is presented for comparison. Fig. 4.10 (a) shows the original signal and Figs. 4.10 (b) and (c) show the signals reconstructed by Lorentzian BP and BPD, respectively. The reconstruction SNR is 115.1 dB for Lorentzian BP and -8.4 dB for BPD. This result illustrates the utility

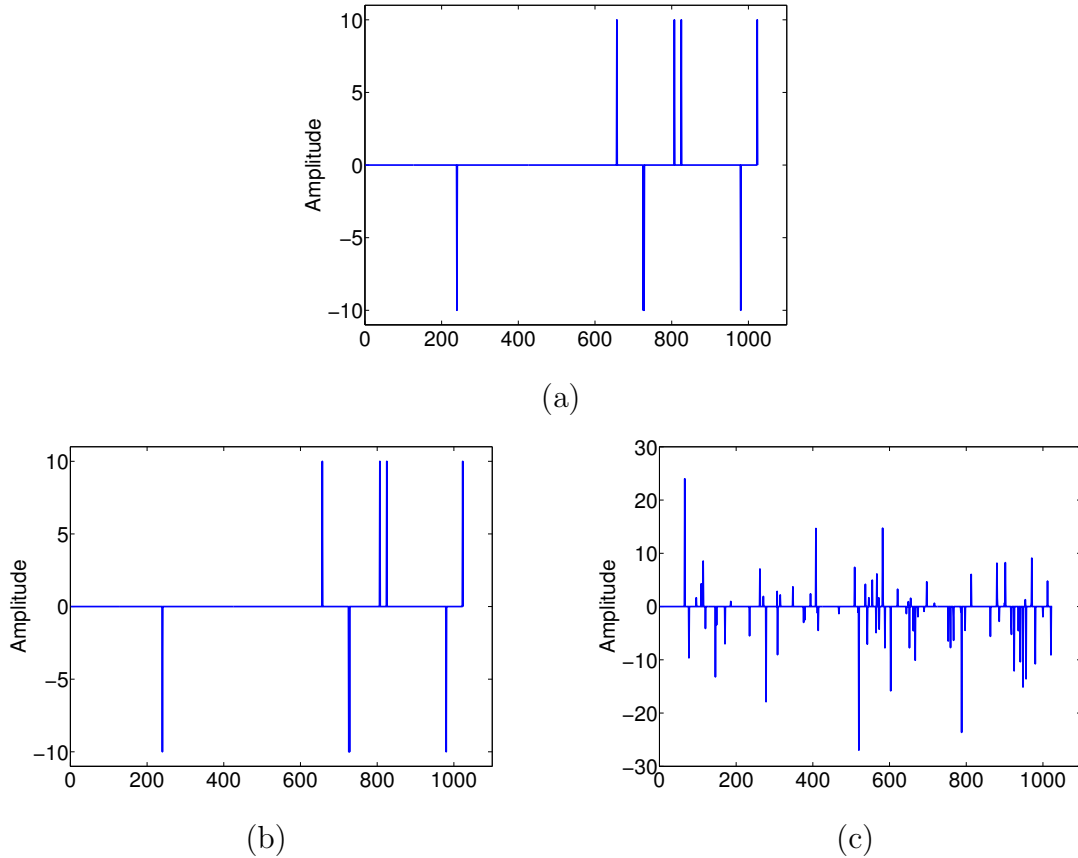


Figure 4.10: Outlier rejection example. (a) Original sparse signal (b) Reconstructed signal using Lorentzian BP SNR=115.1 dB (c) Reconstructed signal using OMP SNR=-8.4 dB.

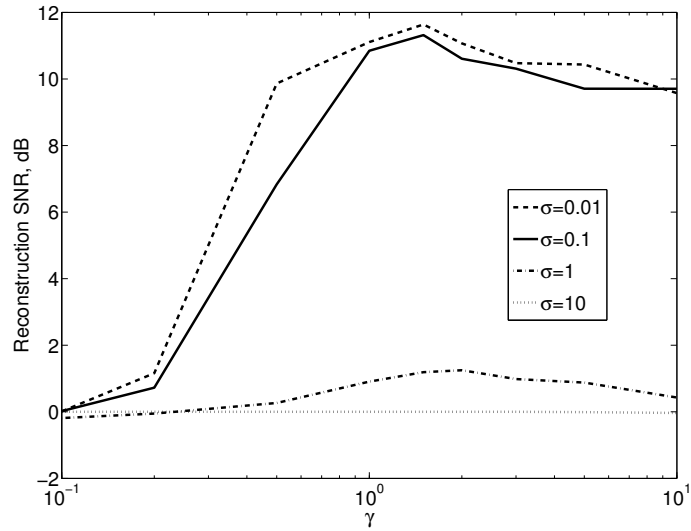
of using a Lorentzian constraint, rather than the commonly employed L_2 constraint on the residual.

In the following experiments we explore the performance of Lorentzian BP as a function of γ , the scale parameter of the Lorentzian norm. Since the Lorentzian metric is derived from Cauchy statistics, the first experiment is performed using standard Cauchy sampling noise for different scale parameters, $\sigma \in \{0.01, 0.1, 1, 10\}$, to evaluate the effect of the noise strength. The second experiment explores the effect of noise impulsiveness; therefore the sampling noise model is α -stable with a fixed scale parameter σ of 0.1 and $\alpha \in \{0.5, 1, 1.5, 2\}$. γ is varied in the interval

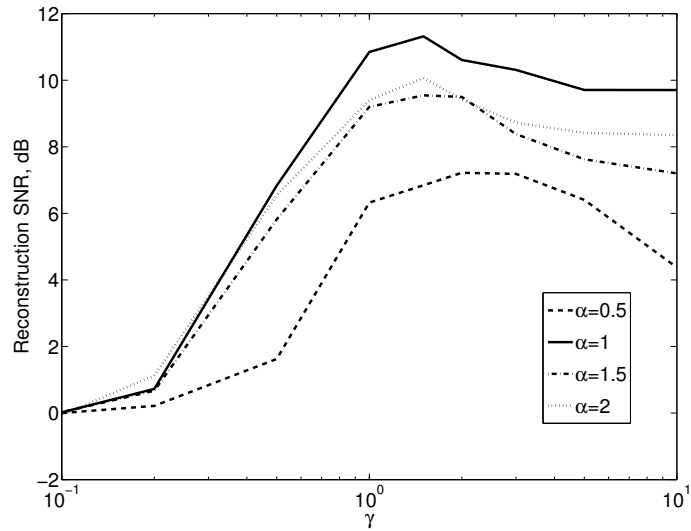
[0.1, 10] for both experiments (since the typical peak amplitude of the uncorrupted measurements for the test signals is 7). Results are summarized in Fig. 4.11 (a) for Cauchy sampling noise and Fig. 4.11 (b) for α -stable sampling noise. In both experiments Lorentzian BP is used without debiasing to explore the effect of γ . The bound for the Lorentzian constraint is set as $\epsilon = 2m \log(1 + \sigma/\gamma)$, where σ is the noise scale parameter. As can be noticed in both cases, the performance peak is in the interval [1.5, 2.5], and is relatively invariant on the noise strength and noise impulsiveness. Moreover, the peak depends more on the projections scale than the noise scale or impulsiveness, validating use of the $\text{MAD}(y)$ as an estimate of γ . In the following we use this estimate of γ for all experiments and set the constraint bound as $\epsilon = 2m \log(1 + \sigma/\gamma)$, with σ being the scale parameter of the sampling noise.

Lorentzian BP is derived from Cauchy statistics; therefore, we present an experiment evaluating the properties of Lorentzian BP in this ideal case. First we show the validity of the error bound in Theorem 6 and the effect of the debiasing operation. In this case the scale parameter was varied from 10^{-3} to 1. The results are presented in Fig. 4.12, showing the L_2 reconstruction error before and after debiasing, along with the theoretical upper bound from (4.20) and the theoretical lower bound for the oracle estimator given in (4.28). The error after debiasing is smaller, as expected, although it has a dramatic increase for $\sigma > 0.1$, when the L_1 optimization does not recover accurately the support of x . An observation of note is that for $2 \cdot 10^{-3} \leq \sigma \leq 10^{-1}$, the reconstruction error of Lorentzian BP is very close to that of the ideal oracle estimator, showing the effectiveness of Lorentzian L_1 minimization to recover the true signal support.

The next set of experiments explore the robustness of Lorentzian BP in different impulsive sampling noises, comparing its performance with OMP and BPD. For OMP and BPD the noise bound is set as $\epsilon = m\sigma^2$, where σ is the scale parameter of



(a)



(b)

Figure 4.11: Reconstruction SNR as a function of γ . (a) Effect of the noise strength, standard Cauchy noise with variable scale parameter σ . (b) Effect of the noise impulsiveness, α -stable noise with variable tail parameter α and fixed scale parameter $\sigma = 0.1$.

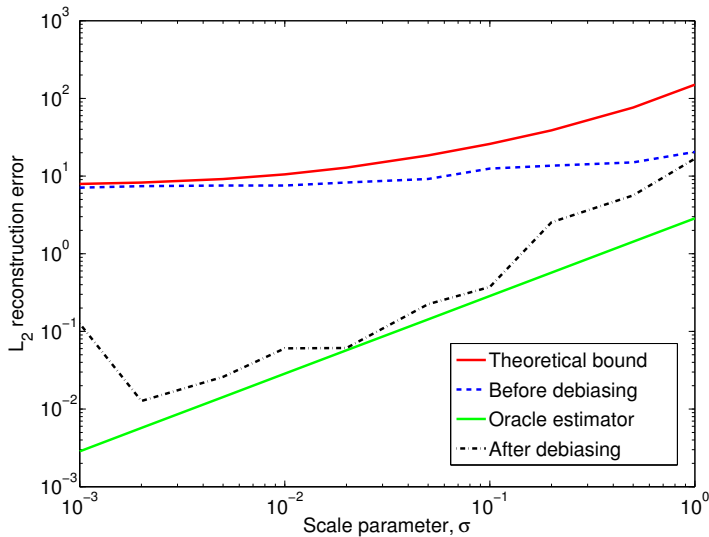


Figure 4.12: L_2 reconstruction error of Lorentzian BP, before and after debiasing for different Cauchy environments. The theoretical upper bound is plotted for comparison.

the corrupting distributions. Fig. 4.13 shows the reconstruction SNR for standard Cauchy sampling noise, with σ varying from 10^{-3} to 10, resulting in a variation of the G-SNR from 28.9 dB to -11.1 dB. The proposed recovery method outperforms both BPD and OMP, which is not surprising since it is optimal under Cauchy statistics. As perhaps a more realistic scenario, we consider contaminated p -Gaussian as the model for the sampling noise, with $\sigma^2 = 10^{-2}$, resulting in an SNR of 18.9 dB when $p = 0$. The amplitude of the outliers is set as $\delta = 10^3$ and p is varied from 10^{-3} to 0.5. The results are shown in Fig. 4.14 (a), which demonstrates that Lorentzian BP significantly outperforms BPD and OMP. Moreover, the Lorentzian BP results are stable over a range of contamination factors p , up to 5% of the measurements making it a desirable method when measurements are lost or erased.

The last experiment explores the behavior of Lorentzian BP in α -stable environments, a noise model particularly instructive, since it contains algebraic-tailed distributions and the light-tailed Gaussian distribution as special cases. The scale

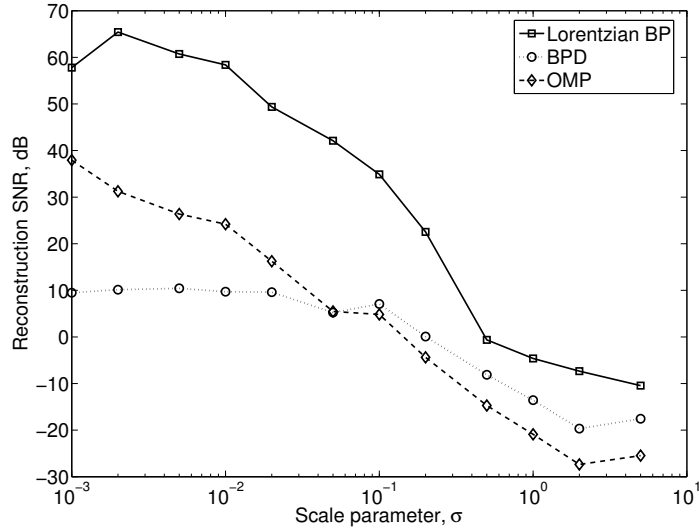
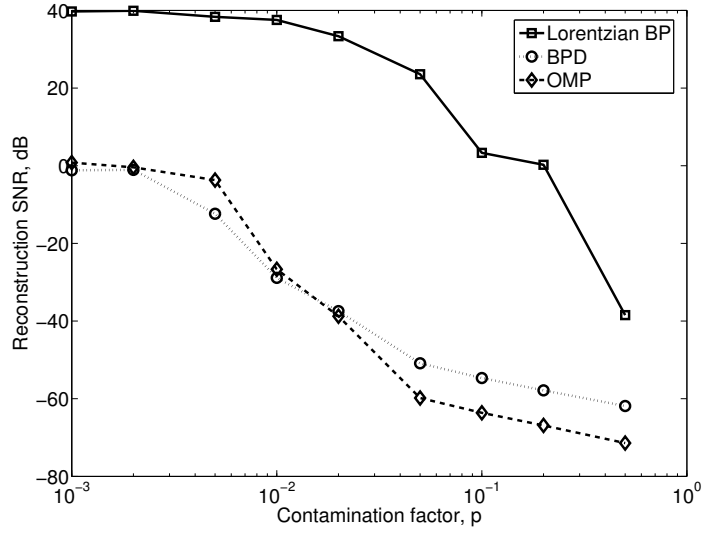


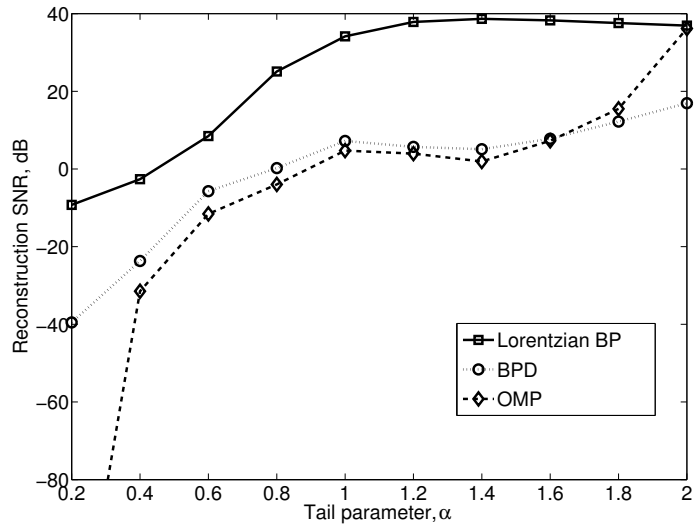
Figure 4.13: Comparison of Lorentzian BP with BPD and OMP in different Cauchy environments. Reconstruction SNR as a function of the scale parameter σ .

parameter of the noise is set as $\sigma = 0.1$ for all cases and the tail parameter, α , is varied from 0.2 to 2, *i.e.* very impulsive to the Gaussian case. The results are summarized in Fig. 4.14 (b) which shows that all methods perform poorly for small values of α , with Lorentzian BP yielding the most acceptable results. Beyond $\alpha = 0.8$, Lorentzian BP produces faithful reconstructions with a SNR greater than 20 dB, and often 30 dB greater than BPD and OMP results. Also of importance is that when $\alpha = 2$ (Gaussian case) the performance of Lorentzian BP is comparable with that of BPD and OMP.

As a final experiment, we evaluate the performance of Lorentzian BP as the number of measurements varies for different levels of impulsiveness. The number of measurements is varied from 16 (twice the sparsity level) to 512 (half the dimension of x). The sampling noise model used is α -stable with four values of α : 0.5, 1, 1.5, 2. The results are summarized in Fig. 4.15, which show that, for $\alpha \in [1, 2]$, Lorentzian



(a)



(b)

Figure 4.14: Comparison of Lorentzian BP with BPD and OMP for impulsive contaminated samples. (a) Contaminated p -Gaussian, $\sigma^2 = 0.01$. R-SNR as a function of the contamination parameter, p . (b) α -S noise, $\sigma = 0.1$. R-SNR as a function of the tail parameter, α .

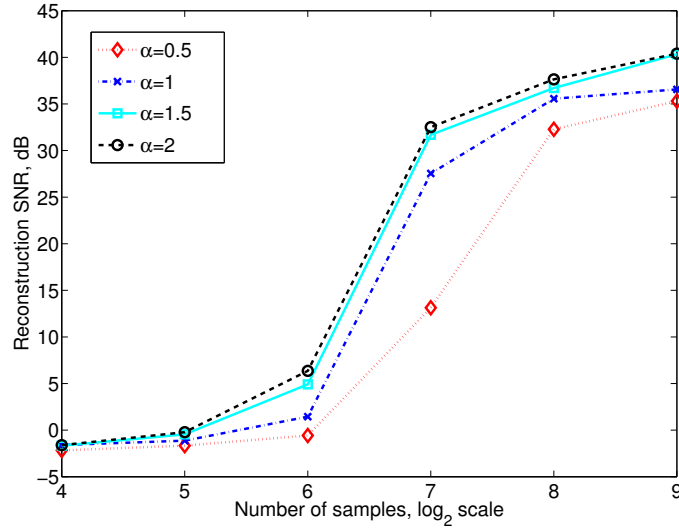


Figure 4.15: Reconstruction SNR as a function of the number of measurements.

BP yields fair reconstructions from 128 samples. However for $\alpha = 0.5$ (most impulsive case of the four), more samples are needed, 256, to yield a fair reconstruction. This result leads to the conclusion that Lorentzian BP can handle heavily corrupted measurements at the expense of requiring more samples, just as LS based methods need more samples for high variance cases.

4.6 Concluding Remarks

This Chapter presents robust sampling and reconstruction methods for sparse signals in impulsive environments. Myriad projections are proposed as sampling operators to address problems with impulsive observation noise. Properties of the proposed sampling function are analyzed, and it is noted that reconstruction performance depends on a linearity parameter, K , which can be adapted to the signal and noise environment. Importantly, myriad projections can be used with standard Gaussian-derived reconstruction algorithms. To address the problem of heavy-tailed sampling noise, Lorentzian basis pursuit is proposed. A reconstruction bound is derived that depends on the noise strength and a tunable parameter of the Lorentzian

norm. Methods to estimate the adjustable parameters in the sampling functions and reconstruction algorithms are proposed, although computation of their optimal values remains an open question. Thus Myriad projections and Lorentzian BP offer a robust framework for CS in impulsive heavy-tailed environments, with performance comparable to existing methods in less demanding light-tailed environments.

Chapter 5

ROBUST BAYESIAN COMPRESSED SENSING USING GENERALIZED CAUCHY MODELS

5.1 Introduction

Compressed sensing shows that a sparse or compressible signal can be reconstructed from a highly incomplete sets of linear measurements [47]. Let $x \in \mathbb{R}^n$ be a sparse signal, and $y = \Phi x$ a set of measurements with Φ an $m \times n$ sensing matrix ($m < n$). The optimal algorithm to recover x from the measurements is

$$\min_x \|x\|_0 \quad \text{subject to} \quad \Phi x = y \quad (5.1)$$

(optimal in the sense that it finds the sparsest vector x such that is consistent with the measurements). Since noise is always present in real data acquisition systems, the acquisition system can be modeled as

$$y = \Phi x + r \quad (5.2)$$

where r represents the sampling noise.

The problem in (5.1) is combinatorial and NP-complete. However, a range of different algorithms have been developed that enable approximate reconstruction of sparse signals from noisy compressive measurements (see [42,44,47,60,65,83,115,134,135,158,161] and references therein). To see a review and comparison of the most relevant algorithms see [134]. The most common approach is to use Basis Pursuit

Denoising (BPD) [47], which uses an unconstrained convex program to estimate a solution of the problem. A family of iterative greedy algorithms (see [134] and references therein) are shown to enjoy a similar approximate reconstruction property, generally with less computational complexity. However, these algorithms require more measurements for exact reconstruction than the L_1 minimization approach.

Recent works show that nonconvex optimization problems can recover a sparse signal with fewer measurements than current geometric methods, while preserving the same reconstruction quality [46, 53, 63, 73, 133, 167]. In [63], the authors replace the L_1 norm in BPD with the L_p norms, for $0 < p < 1$, to approximate the L_0 norm and encourage sparsity in the solution. Candès *et. al* use a re-weighted L_1 minimization approach to find a sparse solution in [46]. The idea is that giving a large weight to small components encourages sparse solutions.

In yet another approach, it is shown that modifying the CS framework to include prior signal knowledge improves the reconstruction results using fewer measurements [22, 90, 104, 119, 164]. Tree structures, for instance in wavelet representations, have also been exploited to introduce prior information in CS signal reconstruction [91, 124], as have Hidden Markov Tree (HMT) models and Markov Random Fields (MRFs) [61, 89]. Baraniuk *et. al* introduced a model-based CS theory that reduces the degrees of freedom of a sparse/compressible signal by permitting only certain configurations of large and zero/small signal coefficients [22, 90]. Similarly, a recovery framework based on a structured union of subspaces is proposed by Eldar and Mishali [97], while source statistics, modeled as stochastic processes, are exploited in [104].

The CS problem can also be treated in a Bayesian framework, where probabilistic priors on the signal coefficients and the corrupting noise are assumed [13–15, 23, 120, 159] and a solution is iteratively constructed. The most common prior utilized in the CS literature is the Laplacian distribution [14, 15], which gives an

statistical justification for the BPD formulation. However, the basic premise in CS is that a small set of coefficients in the signal have larger value than the rest of the coefficients (ideally zero), yielding a very impulsive characterization rather than an exponential-tailed decay behavior. Algebraic-tailed distributions put more mass in very high amplitude values and also in “zero-like” small values, and are therefore more suitable models for sparse coefficients of compressible signals.

In this Chapter, we formulate the CS recovery problem in a Bayesian framework using algebraic-tailed priors from the generalized Cauchy distribution (GCD) family for the signal coefficients and the measurement noise, where the objective is to provide a maximum a posteriori (MAP) signal estimate. This MAP formulation closely resembles L_0 -norm minimization, which features the theoretically lowest bounds on number of measurements required for signal recovery [38]. An iterative reconstruction algorithm is developed from this Bayesian formulation. Simulation results show that GCD priors are a good model for sparse representations. Numerical results also show that the proposed method requires fewer samples than most existing recovery strategies to perform the reconstruction with additional robustness in heavy and light tail noise environments.

The organization of the Chapter is as follows. Section 5.2 gives a brief review of Bayesian modeling and Bayesian CS with exponential priors. In Section 5.3 the CS problem is formulated in a Bayesian framework using GCD priors and an iterative algorithm is proposed to solve the MAP estimation problem. In Section 5.4 the proposed approach is extended to a robust algorithm assuming Cauchy models for the noise. Numerical experiments to evaluate the performance of the proposed algorithms in different environments are presented in Section 5.5. Finally, we close in Section 5.6 with conclusions and future directions.

5.2 Bayesian Modeling and Compressed Sensing

In Bayesian modeling, all unknowns are treated as stochastic quantities with assigned probability distributions. Consider the observation model in (5.2). The unknown signal x is modeled by a *prior* distribution $p(x)$, which represents the *a priori* knowledge about the signal. The observation y is modeled by the likelihood function $p(y|x)$, which is determined in most cases by the noise model. The maximum a posteriori (MAP) estimate of x is given by the solution of the optimization problem

$$\max_{x \in \mathbb{R}^n} p(x|y) = \max_{x \in \mathbb{R}^n} p(y|x)p(x). \quad (5.3)$$

For example, modeling the sampling noise as white Gaussian noise and using a Laplacian prior for x , the MAP estimate of x is equivalent to find the solution of

$$\min_x \|y - \Phi x\|_2^2 + \lambda \|x\|_1, \quad (5.4)$$

which gives statistical justification to the well known LASSO estimator [158].

The statistical behavior of a wide range of process, including DCT and wavelets image coefficients and image pixels difference, can be modeled by the generalized Gaussian distribution (GGD) [4, 24]. The GGD pdf is given by

$$f(x) = \frac{k\alpha}{2\Gamma(1/k)} \exp -(\alpha|x - \theta|)^k \quad (5.5)$$

where $\Gamma(\cdot)$ is the gamma function $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ and θ is the location parameter. In this form, α is an inverse scale parameter and $k > 0$, sometimes called the shape parameter, controls the tail decay rate. The GGD model contains the Laplacian and Gaussian distributions as special cases, *i.e.*, for $k = 1$ and $k = 2$, respectively. The Laplacian is the most common prior utilized in the CS literature [14, 15, 120]. However, [13] utilizes the GGD to model the sparse signal coefficients and provide a statistical justification to the use of non convex priors (L_p

norms with $0 < p < 1$) in CS reconstruction.

Even though the GGD has been successfully utilized in Bayesian CS, the basic premise of compressible models is that a small set of coefficients in the signal have larger value than the rest of the coefficients (ideally zero), yielding a very impulsive characterization rather than an exponential-tailed decay behavior. Algebraic-tailed distributions put more mass in very high amplitude values and also in “zero-like” small values, and are therefore more suitable models for sparse coefficients of compressible signals (see also [111]). Therefore, in the next section we propose a Bayesian CS theory based on the generalized Cauchy family.

5.3 Bayesian Compressed Sensing with Generalized Cauchy Priors

5.3.1 MAP estimation with generalized Cauchy priors

Of interest here is the development of a sparse reconstruction strategy using a Bayesian framework. To encourage sparsity in the solution, we propose the use of GC priors for the signal model. The GC family of distributions possesses heavier tails than the Laplacian, thus yielding more impulsive (sparser) signal models and intuitively lowering the number of samples to perform the reconstruction.

Recall that the PDF of the GCD is given by

$$f(z) = a\delta(\delta^p + |z|^p)^{-\frac{2}{p}} \quad (5.6)$$

with $a = p\Gamma(2/p)/2(\Gamma(1/p))^2$. In this representation, δ is the scale parameter and p is the tail constant. The GCD has been used to model many impulsive processes in real life (see Chapter 3).

We model the sampling noise as independent, zero mean, Gaussian distributed samples with variance σ^2 . Using the observation model in (5.2) the likelihood function becomes

$$p(y|x; \sigma) = \mathcal{N}(\Phi x, \Sigma), \quad \Sigma = \sigma^2 I. \quad (5.7)$$

Assuming the signal x (or coefficients in a sparse basis) are independent GC distributed samples yields the following prior

$$p(x|\delta, p) = (a\delta)^n \prod_{i=1}^n (\delta^p + |x_i|^p)^{-2/p} \quad (5.8)$$

Since $p(x|y; \sigma, \delta, p) \propto p(y|x; \sigma)p(x|\delta, p)$, the MAP estimate, assuming σ , δ and p known, is

$$\hat{x} = \arg \min_x \frac{1}{2} \|y - \Phi x\|_2^2 + \lambda \|x\|_{LL_{p,\delta}} \quad (5.9)$$

where $\lambda = 2\sigma^2$.

One remark to make is that the LL_1 norm has been previously used to approximate the L_0 norm but without making a statistical connection to the signal model. The re-weighted L_1 approach proposed in [46] is equivalent to finding a solution for the first order approximation of the problem

$$\min_x \|x\|_{LL_{1,\delta}}, \text{ s.t. } \|y - \Phi x\|_2 \leq \epsilon, \quad (5.10)$$

using a decreasing sequence for δ .

5.3.2 Algorithm formulation

In this work, instead of directly minimizing (5.9), we develop a fixed point search to find a sparse solution. The fixed point algorithm is based on first order optimality conditions and is inspired in the robust statistics literature [118].

Let x^* be a stationary point of (5.9), then the first order optimality condition is

$$\Phi^T \Phi x^* - \Phi^T y + \lambda \nabla_x \|x^*\|_{LL_p, \delta} = 0. \quad (5.11)$$

Noting that the gradient $\nabla_x \|x^*\|_{LL_p, \delta}$, can be expressed as

$$\nabla_x \|x^*\|_{LL_p, \delta} = W(x^*)x^*, \quad (5.12)$$

where $W(x)$ is a diagonal matrix with diagonal elements given by

$$[W(x)]_{ii} = [(\delta^p + |x_i|^p)|x_i|^{2-p}]^{-1}, \quad (5.13)$$

the first order optimality condition, (5.11), is equivalent to

$$\Phi^T \Phi x^* - \Phi^T y + \lambda W(x^*)x^* = 0. \quad (5.14)$$

Solving for x^* we find the fixed point function

$$\begin{aligned} x^* &= [\Phi^T \Phi + \lambda W(x^*)]^{-1} \Phi^T y \\ &= W^{-1}(x^*) \Phi^T [\Phi W^{-1}(x^*) \Phi^T + \lambda I]^{-1} y. \end{aligned} \quad (5.15)$$

The fixed point search uses the solution at previous iteration as input to update the solution. The estimate at iteration time $t + 1$ is given by

$$\hat{x}_{t+1} = W^{-1}(\hat{x}_t) \Phi^T [\Phi W^{-1}(\hat{x}_t) \Phi^T + \lambda I]^{-1} y. \quad (5.16)$$

The fixed point algorithm turns out to be a reweighted least squares recursion [167], which iteratively finds a solution and updates the weight matrix using (5.13). This

estimation reweighted least squares procedure might also be seen as a convex bounding type II variational method for Bayesian estimation.

As in other robust regression problems, the estimate in (5.9) is scale dependent (δ in the GC prior formulation). In fact, δ controls the sparsity of the solution and in the limiting case when $\delta \rightarrow 0$ the solution of (5.9) is equivalent to the L_0 norm solution [46,167]. To address this problem we propose to jointly estimate δ and x at each iteration similar to joint scale-location estimates [48,118]. We use a Type II maximum likelihood approach [140], which is essentially an EM algorithm [129], where we estimate the signal x and then we estimate the prior parameters δ, p from the estimated x . The algorithm consists of alternately updating the prior parameters and updating the signal estimate. We describe the resulting algorithm in the following.

A fast way to estimate δ from x is using order statistics (although more elaborate estimates can be used as in [48,51]). Let X be a GC distributed random variable with zero location and scale parameter δ and denote the r -th quartile of X as $Q_{(r)}$. The interquartile distance is $Q_{(3)} - Q_{(1)} = 2\delta$, thus, a fast estimate of δ is half the interquartile distance of the samples x . Let $Q_{(r)}^t$ denote the r -th quartile of the estimate \hat{x}_t at time t , then the estimate of δ at iteration time t is given by

$$\hat{\delta}_t = 0.5(Q_{(3)}^t - Q_{(1)}^t). \quad (5.17)$$

To estimate p we follow a maximum likelihood approach and maximize the likelihood function given x and δ . The estimate of p at time t is given by

$$\hat{p}_t = \max_{p \in (0,2]} p(x|\delta, p). \quad (5.18)$$

Experimental results show that selecting the tail parameter p from a discrete set of values will not degrade the performance of the reconstruction. We use the set

$\Gamma = \{0.5, 1, 1.5, 2\}$ as a search space for p , thus the estimate of p is given by

$$\hat{p}_t = \max_{p \in \Gamma} p(x|\delta, p). \quad (5.19)$$

To summarize, the final algorithm is depicted in Algorithm 5, where J is the maximum number of iterations and γ is a tolerance parameter for the error between subsequent solutions. To prevent numerical instabilities we pre-define a minimum value for δ denoted as δ_{min} . We start the recursion with the LS solution ($W = I$) and we also assume a known noise variance, σ^2 (recall $\lambda = 2\sigma^2$). The resulting algorithm is coined Generalized Cauchy Bayesian compressed sensing (GCBCS).

Algorithm 5 GCBCS-I

Require: λ , δ_{min} , γ and J .

- 1: Initialize $t = 0$ and $\hat{x}_0 = \Phi^T(\Phi\Phi^T + \lambda I)^{-1}y$.
 - 2: **while** $\|\hat{x}_t - \hat{x}_{t-1}\|_2 > \gamma$ or $t < J$ **do**
 - 3: Update $\hat{\delta}_t$ and p .
 - 4: Update the matrix W .
 - 5: Compute \hat{x}_{t+1} as in equation (5.16).
 - 6: $t \leftarrow t + 1$
 - 7: **end while**
 - 8: **return** \hat{x}
-

As mentioned in the last section the reweighted L_1 approach of [46] and GCBCS with $p = 1$ minimize the same objective. Moreover, the reweighted L_1 may require fewer iterations to converge, but the computational cost of one iteration of GCBCS is substantially lower than the computational cost of an iteration of reweighted L_1 , thereby resulting in a faster algorithm.

5.4 Robust Bayesian Compressed Sensing with Generalized Cauchy Models

5.4.1 MAP estimation with generalized Cauchy priors and noise models

The Bayesian formulation above assumes the measurements are corrupted by Gaussian noise, therefore limiting the robustness of the derived estimators in impulsive sampling noise. To address this problem, we model the noise as zero location i.i.d. GCD samples, with tail parameter q and scale parameter σ . The likelihood function of the observations becomes:

$$p(y|x, \sigma, q) = (a\sigma)^m \prod_{i=1}^m (\sigma^q + |y_i - \theta_i|^q)^{-2/q} \quad (5.20)$$

with location vector $\theta = \Phi x$. Assuming a GC prior with tail parameter p the MAP estimate is given by:

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} \|y - \Phi x\|_{LL_q, \sigma} + 2\|x\|_{LL_p, \delta}. \quad (5.21)$$

As shown in Chapter 4 the Lorentzian norm (derived norm for the standard Cauchy statistics case) possess several desirable properties to be used as a robust fidelity measure. Therefore if we assume a Cauchy distribution for the noise ($q = 2$), the MAP estimator, for σ , δ and p known, is:

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} \|y - \Phi x\|_{LL_2, \sigma} + 2\|x\|_{LL_p, \delta}. \quad (5.22)$$

The estimator defined in (5.22) results in a Lorentzian fidelity term that, as shown in Chapter 4, offers robustness against heavy-tailed and light-tailed noise models. The proposed Bayesian framework yields a simple, yet powerful, family of estimators capable of recovering sparse or compressible signals from samples corrupted by impulsive noise and with fewer samples than traditional recovery strategies.

5.4.2 Fixed point algorithm

Let x^* be a stationary point of (5.22), then the first order optimality condition is

$$\nabla_x \|y - \Phi x\|_{LL_2, \sigma} + 2\nabla_x \|x^*\|_{LL_p, \delta} = 0. \quad (5.23)$$

We know that the gradient $\nabla_x \|x^*\|_{LL_p, \delta}$ can be expressed as

$$\nabla_x \|x^*\|_{LL_p, \delta} = W(x^*)x^*. \quad (5.24)$$

We can use a similar representation for $\nabla_x \|y - \Phi x\|_{LL_2, \sigma}$. Denote ϕ_i as the i -th row vector of Φ . The gradient can be written as

$$\nabla_x \|y - \Phi x\|_{LL_2, \sigma} = \Phi^T H(y - \Phi x^*) \quad (5.25)$$

where H is an $m \times m$ diagonal matrix with each element on the diagonal defined as

$$[H]_{i,i} = \frac{\sigma^2}{\sigma^2 + (y_i - \phi_i^T x^*)^2}, \quad i = 1, \dots, m. \quad (5.26)$$

the first order optimality condition, (5.23), is then equivalent to

$$\Phi^T H \Phi x^* - \Phi^T H y + \lambda W(x^*)x^* = 0. \quad (5.27)$$

Solving for x^* we find the fixed point function

$$\begin{aligned} x^* &= [\Phi^T H \Phi + \lambda W(x^*)]^{-1} \Phi^T H y \\ &= W^{-1}(x^*) \Phi^T [H \Phi W^{-1}(x^*) \Phi^T + 2I]^{-1} H y. \end{aligned} \quad (5.28)$$

The fixed point search uses the solution at previous iteration as input to

update the solution. The estimate at iteration time $t + 1$ is given by

$$\hat{x}_{t+1} = W^{-1}(\hat{x}_t)\Phi^T[H\Phi W^{-1}(\hat{x}_t)\Phi^T + 2I]^{-1}Hy. \quad (5.29)$$

The performance of the GCBCS algorithm depends on the scale parameter σ of the Lorentzian norm and the step size. In [57] is observed that setting σ as half the sample range of y , $(y_{(1)} - y_{(0)})/2$ (where $y_{(q)}$ denotes the q -th quantile of y), often makes the Lorentzian norm a fair approximation to the L_2 norm. Therefore, the optimal value of σ should be $(y'_{(1)} - y'_{(0)})/2$, where $y' = \Phi x_0$ is the uncorrupted measurement vector. Since the uncorrupted measurements are unknown, we propose to estimate the scale parameter as

$$\hat{\sigma} = \frac{y_{(0.875)} - y_{(0.125)}}{2}. \quad (5.30)$$

This value of σ considers implicitly a measurement vector with 25% of the samples corrupted by outliers and 75% well behaved. Experimental results show that this estimate leads to good performance in both Gaussian and impulsive environments. The parameters δ and p are estimated using the same maximum likelihood approach used in Section 5.3.2.

Algorithm 6 GCBCS-II

Require: δ_{min} , γ and J .

- 1: Estimate $\hat{\sigma}$
 - 2: Initialize $t = 0$ and $\hat{x}_0 = \Phi^T(\Phi\Phi^T + \lambda I)^{-1}Hy$.
 - 3: **while** $\|\hat{x}_t - \hat{x}_{t-1}\|_2 > \gamma$ or $t < J$ **do**
 - 4: Update $\hat{\delta}_t$ and \hat{p} .
 - 5: Update H and W .
 - 6: Compute \hat{x}_{t+1} as in equation (5.29).
 - 7: $t \leftarrow t + 1$
 - 8: **end while**
 - 9: **return** \hat{x}
-

Table 5.1: Comparison of reconstruction quality between known δ and estimated δ MBCS. Meridian distributed signals, $n = 1000$, $m = 200$. R-SNR (dB).

	$\delta = 10^{-3}$	$\delta = 10^{-2}$	$\delta = 10^{-1}$
Known δ	9.91	21.5	30.69
Estimated δ	8.16	17.58	24.98

5.5 Experimental Results

5.5.1 Noiseless and light-tailed noise cases

In this section we present numerical experiments that illustrate the effectiveness of MBCS for sparse and compressible signal reconstruction. For all experiments we use random Gaussian measurements matrices with normalized columns and $\delta_{min} = 10^{-8}$ in the algorithm.

The first experiment shows the validity of the joint estimation approach of MBCS. Meridian distributed signals with length $n = 1000$ and $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ are used. The signals are sampled taking $m = 200$ measurements and zero mean Gaussian distributed sampling noise with variance $\sigma^2 = 10^{-2}$ is added. Table 5.1 shows the average reconstruction SNR (R-SNR) for 200 repetitions. The performance loss is of 6 dB approximately in the worst case, but fully automated MBCS still yields a good reconstruction.

The next set of experiments compare GCBCS with current reconstruction strategies for noiseless samples and noisy samples. The algorithms used for comparison are L_1 minimization [47], re-weighted L_1 minimization [46], RWLS to approach L_p [63], and CoSaMP [134]. We use k -sparse signals (k nonzero coefficients) of length $n = 1000$, in which the amplitudes of the nonzero coefficients are Gaussian distributed with zero mean and standard deviation $\sigma_x = 10$. Each experiment is averaged over 200 repetitions.

The first experiment compares GCBCS in a noiseless setting for different sparsity levels, fixing $m = 200$. We use the probability of exact reconstruction as a

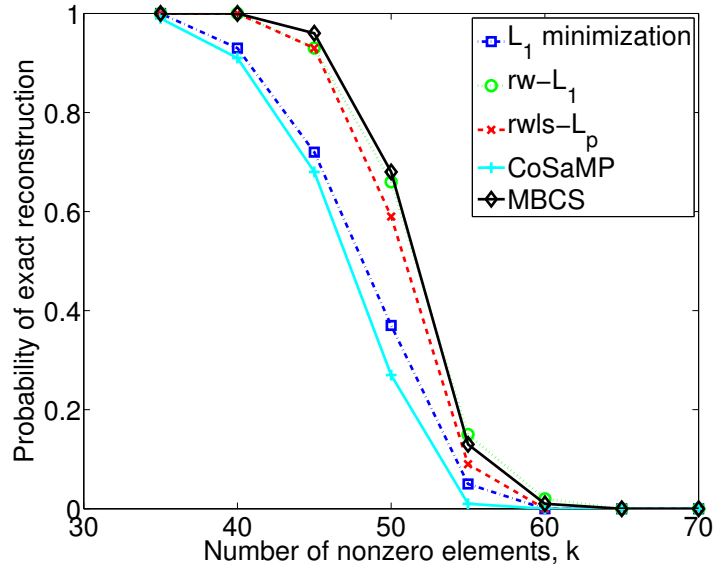


Figure 5.1: Probability of successful recovery as function of the sparsity level k (noiseless case). $m = 200$.

measure of performance, where a reconstruction is considered exact if $\|\hat{x} - x\|_\infty \leq 10^{-4}$. The results are shown in Fig. 5.1. Results show that GCBCS outperforms CoSaMP and L_1 minimization (giving larger probability of success for larger values of k) and yielding a slightly better performance than L_p minimization. It is of notice that GCBCS has similar performance to reweighted L_1 , since they are minimizing the same objective, but with a different approach.

The second experiment compares GCBCS in the noisy case, varying the number of samples (m) and fixing $k = 10$. The sampling noise is Gaussian distributed with variance $\sigma^2 = 10^{-2}$. The R-SNR is used as the performance metric. Results are presented in Fig. 5.2. In the noisy case GCBCS outperforms all other reconstruction strategies, yielding a larger R-SNR for fewer samples with a good approximation for 60 samples and above. Moreover, the R-SNR of GCBCS is better than reweighted L_1 minimization. An explanation for this is that L_1 minimization methods suffer from bias problems needing a de-biasing step after the solution is found (see [57]

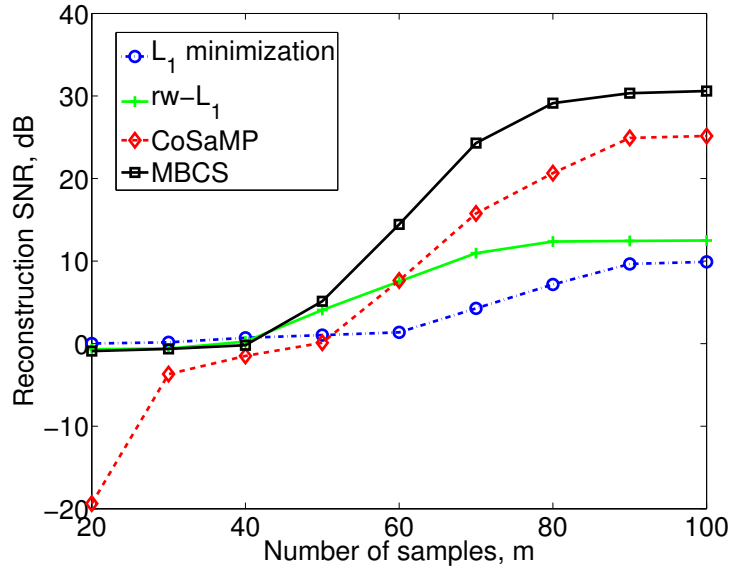


Figure 5.2: Reconstruction SNR as function of the number of samples m (Gaussian sampling noise, $\sigma^2 = 10^{-2}$). Gaussian distributed non-zero coefficients, $\sigma_x = 10$ and $k = 10$.

and references therein) to achieve a similar performance.

The next experiment illustrates the performance of GCBCS for real compressible signals. ECG signals are utilized due to the structure of their sparse decompositions. Experiments are carried out over 10-min long leads extracted from records 100, 101, 102, 103, 107, 109, 111, 115, 117, 118 and 119 from the MIT-BIH Arrhythmia Database (see [30] and references therein). Cosine modulated filter banks are used to determine a sparse representation of the signal [30]. A sparse signal approximation is determined by processing 1024 samples of ECG data, setting the number of channels, M , to 16. R-SNR is used as the performance metric. Results are presented in Figure 5.3. In the compressible case GCBCS outperforms all other reconstruction algorithms, yielding a larger R-SNR for fewer samples with a good approximation obtained from 300 samples (R-SNR greater than 20 dB). One remark is that for $m < 200$ GCBCS reconstruction results are worse than L_1 minimization,

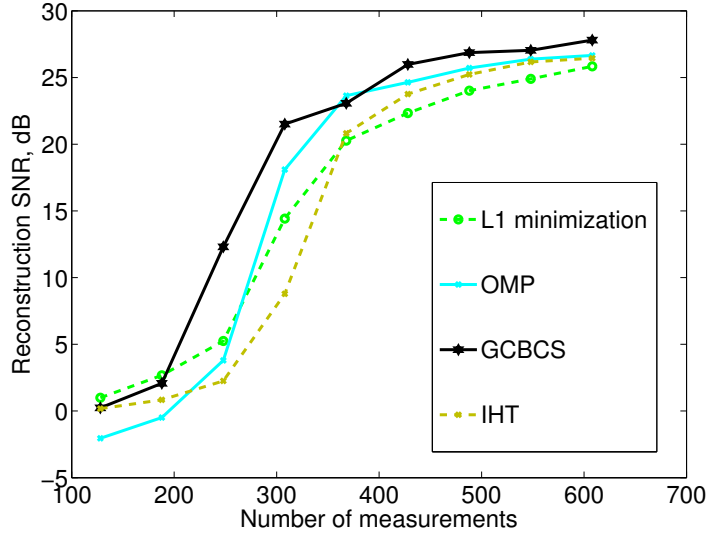


Figure 5.3: Reconstruction SNR as function of the number of samples m . ECG signals using CMFB, $M = 16$ and $n = 1024$.

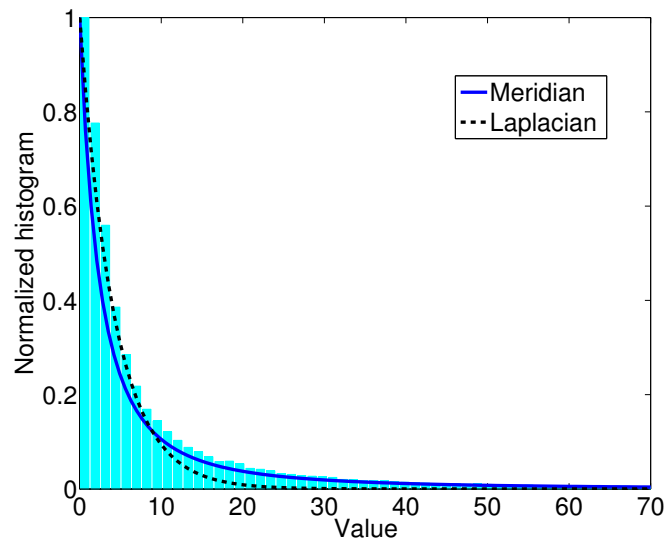
losing 1 dB in the R-SNR.

As an illustrative example for image models, we present an example utilizing a 256×256 image. We use a Daubechies db8 wavelets as sparse basis. Fig. 5.4 (a) shows the boats image and Fig. 5.4 (b) shows a zoom of the normalized histogram of its coefficients along with a plot of meridian and Laplacian distributions. It can be noticed that the meridian is a better fit for the tails of the coefficient distribution.

To show the effectiveness of GCBCS to model and recover real images we sample and reconstruct a set of 10 256×256 ($n = 65536$) standard test images. We employ a partial random Hadamard ensemble to sample the images and the number of measurements, m , is varied from 4000 to 40000. We also use a Daubechies db8 wavelets as sparsity basis. We compare CGBCS against Laplacian Bayesian compressed sensing (LBCS) [14,15] and the iterative hard thresholding (IHT) algorithm. We use the PSNR as performance metrics for this experiment. The results are averaged over the 10 different images and over 100 different realizations of the measurement matrix for each image. The results are shown in Fig. 5.5 and it can



(a)



(b)

Figure 5.4: Image model example. (a) Original image, (b) Wavelet coefficient histogram with Laplacian distribution fit (dashed) and Meridian distribution fit (blue).

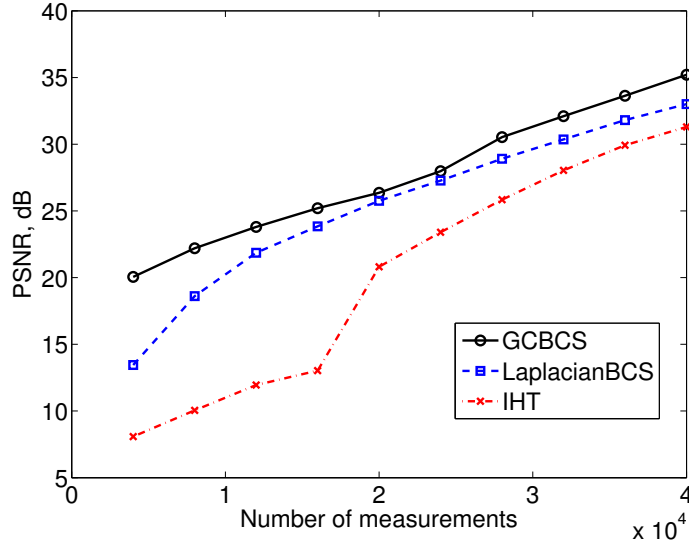


Figure 5.5: PSNR as function of the number of samples m . Average results on 10 256×256 images.

be seen that GCBCS outperformed both LBCS and IHT. Moreover, the PSNR improvement over LBCS and IHT is more notorious in the highly undersampled regime $m \in [8000, 12000]$, which confirms that the generalized Cauchy models are better priors for sparse and compressible signals.

Fig. 5.6 shows a realization of this experiment for the Lena image. The top row shows the recovered images for $m = 8000$ obtained by LBCS (left), PSNR=18.61 dB and by GCBCS (right), PSNR=23.81 dB. The middle row shows the recovered images for $m = 20000$ obtained by LBCS (left), PSNR=25.56 dB and by GCBCS (right), PSNR=26.36 dB. The bottom row shows the recovered images for $m = 32000$ obtained by LBCS (left), PSNR=30.36 dB and by GCBCS (right), PSNR=32.10 dB. As mention above, the performance improvement of generalized Cauchy priors over Laplacian priors is more noticeable for the highly undersampled ($m = 8000$) case. In this case the reconstructed image from the LBCS approach loses all the face details. On the other hand, GCBCS preserves most of the face details.



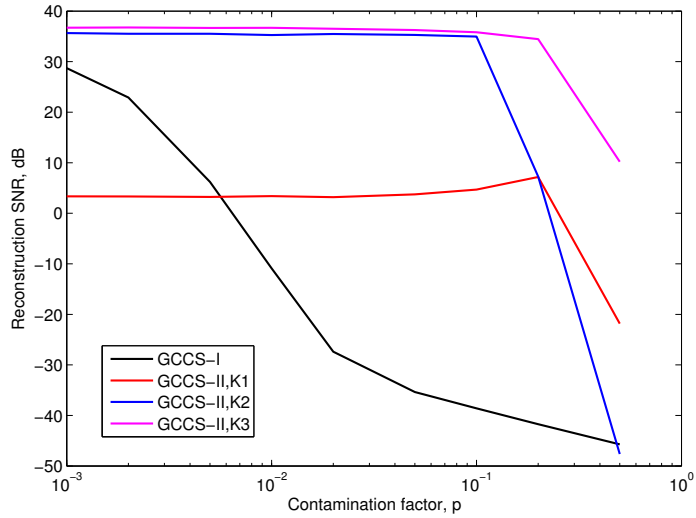
Figure 5.6: Image reconstruction example with Lena. Top row: $m = 8000$. LBCS (left), PSNR=18.61 dB and GCBCS (right), PSNR=23.81 dB. Middle row: $m = 20000$. LBCS (left), PSNR=25.56 dB and GCBCS (right), PSNR=26.36 dB. Bottom row $m = 32000$. LBCS (left), PSNR=30.36 dB and GCBCS (right), PSNR=32.10 dB.

5.5.2 Heavy-tailed noise

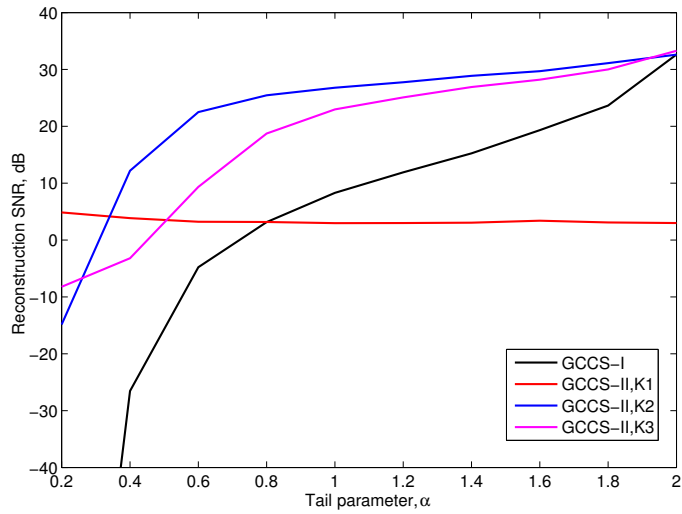
Numerical experiments that illustrate the effectiveness of the GCBCS-II in impulsive environments algorithm are presented in this section. All experiments utilize synthetic s -sparse signals in a Hadamard basis, with $s = 8$ and $n = 1024$. The nonzero coefficients have equal amplitude, equiprobable sign, randomly chosen position, and average power fixed to 0.78. Gaussian sensing matrices are employed with $m = 128$. One thousand repetitions of each experiment are averaged and reconstruction SNR is used as the performance measure. We compare the GCBCS-II algorithm to the Gaussian noise derived GCBCS-I.

To test the robustness of the methods, we use two noise models: α -stable distributed noise and Gaussian noise plus gross sparse errors. The Gaussian noise plus gross sparse errors model is referred to as contaminated p -Gaussian noise for the remainder of the paper, as p represents the amount of gross error contamination. To validate the estimate of σ discussed in Section 5.4.2 we make a comparison between the performance of GCBCS equipped with the median absolute (MAD) of y as an estimate σ , denoted as GCBCS-II,K1, the proposed estimator of σ , denoted as GCBCS-II,K2 and the optimal σ , denoted as GCBCS-II,K3. The optimal σ is set as half the sample range of the clean measurements.

For the first experiment we consider a mixed noise environment, using contaminated p -Gaussian noise. We set the Gaussian component variance to $\sigma^2 = 10^{-2}$, resulting in an SNR of 18.9321 dB when $p = 0$. The amplitude of the outliers is set as $\delta = 10^3$ and p is varied from 10^{-3} to 0.5. The results are shown in Fig. 5.7 (a). The results demonstrate that GCBCS-II outperforms GCBCS-I. Moreover, the results also demonstrate the validity of the estimated σ . Although the reconstruction quality achieved by GCBCS-II,K2 is lower than that achieved GCBCS-II,K3, the SNR of GCBCS-II,K2 is greater than 20 db for a broad range of contamination factors p , including contaminations up to 1% of the measurements.



(a)



(b)

Figure 5.7: Comparison of GCBCS for impulsive contaminated samples. (a) Contaminated p -Gaussian, $\sigma^2 = 0.01$. R-SNR as a function of the contamination parameter, p . (b) α -stable noise, $\sigma = 0.1$. R-SNR as a function of the tail parameter, α .

The second experiment explores the behavior of GCBCS in very impulsive environments. We use this time with α -Stable sampling noise. The scale parameter of the noise is set as $\sigma_n = 0.1$ for all cases and the tail parameter, α , is varied from 0.2 to 2, *i.e.*, very impulsive to the Gaussian case, Fig. 5.7 (b). For small values of α , all methods perform poorly, with GCBCS-II,K2 yielding the most acceptable results. Beyond $\alpha = 0.6$, GCBCS-II,K2 produces faithful reconstructions with a SNR greater than 20 dB. It is of notice that when $\alpha = 2$ (Gaussian case) the performance of GCBCS-II,K2 is comparable with that of GCBCS-I, which is Gaussian derived. Also of notice is that the SNRs achieved by GCBCS-II,K2 is better than the one achieved by GCBCS-II,K3 and GCBCS-II,K1.

For the last experiment, we evaluate the performance of GCBCS-II as the number of measurements varies for different levels of impulsiveness. The number of measurements is varied from 16 (twice the sparsity level) to 512 (half the dimension of x_0). The sampling noise model used is α -stable with four values of α : 0.5, 1, 1.5, 2. The results are summarized in Fig. 5.8, which show that, for $\alpha \in [1, 2]$, GCBCS-II yields fair reconstructions from 64 samples. However for $\alpha = 0.5$ (most impulsive case of the four), more samples are needed, 256, to yield a fair reconstruction. Results of GCBCS-II with Gaussian noise ($\alpha = 2$) are also included for comparison. It is of notice that the performance of GCBCS-II is comparable to that of GCBCS-I for the Gaussian case. Another interesting conclusion is that the reconstruction quality of GCBCS is better than that obtained for the Lorentzian BP approach presented in Chapter 4 with the same number of measurements.

5.6 Concluding Remarks

In this Chapter, we formulate the CS recovery problem in a Bayesian framework using algebraic-tailed priors from the GCD family for the signal coefficients and the measurement noise. We show that algebraic-tailed impulsive distributions are more suitable models for sparse or compressible signals a conclusion also shown

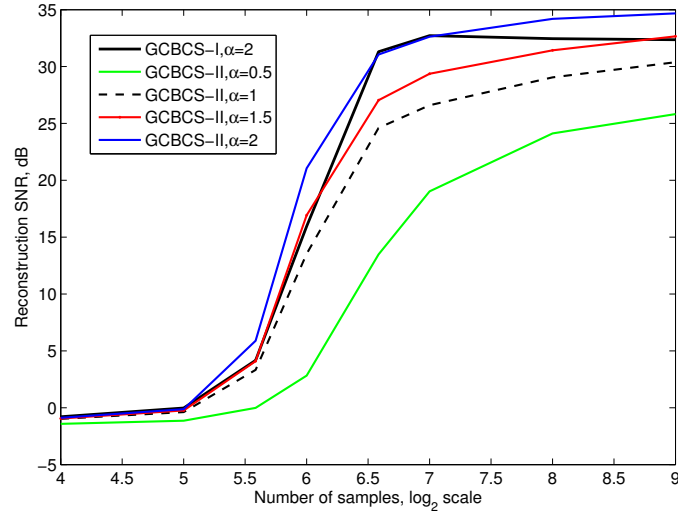


Figure 5.8: Performance of GCBCS-II as the number of measurements varies for synthetic sparse signals. Reconstruction SNR as a function of the number of measurements.

in [111]. An iterative reconstruction algorithm, referred to as GCBCS, is developed from this Bayesian formulation. Simulation results show that the proposed method requires fewer samples than most existing reconstruction algorithms for compressed sensing, thereby validating the use of GCD priors for sparse reconstruction problems. The proposed Bayesian yields comparable performance with state of the art algorithms in light-tailed noise environments while having substantial performance improvements in heavy-tailed environments.

Chapter 6

LORENTZIAN ITERATIVE HARD THRESHOLDING: ROBUST COMPRESSED SENSING WITH PRIOR INFORMATION

6.1 Introduction

Compressed sensing (CS) demonstrates that a sparse, or compressible, signal can be acquired using a low rate acquisition process that projects the signal onto a small set of vectors incoherent with the sparsity basis [47]. There are several reconstruction methods that yield perfect or approximate reconstruction proposed in the literature (see [34, 47, 134] and references therein). To see a review and comparison of the most relevant algorithms see [134]. Since noise is always present in practical acquisition systems, a range of different algorithms and methods have been developed that enable approximate reconstruction of sparse signals from noisy compressive measurements [34, 47, 134]. Most such algorithms provide bounds for the L_2 reconstruction error based on the assumption that the corrupting noise is Gaussian, bounded, or, at a minimum, has finite variance. In contrast to the typical Gaussian assumption, heavy-tailed processes exhibit very large, or infinite, variance. Existing reconstruction algorithms operating on such processes yield estimates far from the desired original signal.

Recent works have begun to address the reconstruction of sparse signals from measurements corrupted by impulsive processes [5, 57, 125, 142]. Laska *et al.* assume a sparse error and estimate both signal and error at the same stage [125]. Carrillo

et al. propose a reconstruction approach based on robust statistics theory [57]. The proposed non-convex program seeks a solution that minimizes the L_1 norm subject to a nonlinear constraint based on the Lorentzian norm. Following this line of thought, this approach is extended in [5] to develop an iterative algorithm to solve a Lorentzian L_0 -regularized cost function using iterative weighted myriad filters. A similar approach is used in [142] by solving an L_0 -regularized least absolute deviation regression problem, yielding an iterative weighted median algorithm. Even though these approaches provide a robust CS framework in heavy-tailed environments, numerical algorithms to solve the proposed optimization problem are slow and complex, especially as the dimension of the problem grows.

Recent results in CS show that modifying the recovery framework to include prior knowledge of the support improves the reconstruction results using fewer measurements [119, 164]. Vaswani *et al.* assume that part of the signal support is known *a priori* and the problem is recast as finding the unknown support. The remainder of the signal (unknown support) is a sparser signal than the original, thereby requiring fewer samples to yield an accurate reconstruction [164]. Although the modified CS approach in [164] needs fewer samples to recover a signal, it employs a modified version of basis pursuit (BP) [47] to perform the reconstruction. The computational cost of solving the convex problem posed by BP can be high for large scale problems. Therefore, in [58] we proposed to extend the ideas of modified CS to iterative approaches like greedy algorithms [134] and iterative re-weighted least squares methods [63]. These algorithms construct an estimate of the signal at each iteration, and are thereby amenable to incorporation of *a priori* support information (1) as an initial condition or (2) at each iteration. Although the aforementioned methods are more efficient than BP, in terms of computational cost, a disadvantage of these methods is the need to invert a linear system at each iteration.

In this Chapter we propose a Lorentzian based iterative hard thresholding

(IHT) algorithm and a simple modification to incorporate prior signal information in the recovery process. Specifically, we study the case of CS with partially known support. The IHT algorithm is a simple iterative method that does not require matrix inversion and provides near-optimal error guarantees [32,33]. Hard thresholding algorithms have been previously used in image denoising [26] and sparse representations [72,94,117]. All of these methods are particular instances of a more general class of iterative thresholding algorithms [100,138,155]. A good general overview of iterative thresholding methods is given in [95]. Related convergence results can be found in [69].

The proposed algorithm is a fast method with computational load comparable to the least squares (LS) based IHT, whilst having the advantage of robustness against heavy-tailed impulsive noise. Sufficient conditions for stability are studied and a reconstruction error bound is derived. We also derive sufficient conditions for stable sparse signal recovery with partially known support. Theoretical analysis shows that including prior support information relaxes the conditions for successful reconstruction. Simulations results demonstrate that the Lorentzian-based IHT algorithm significantly outperform commonly employed sparse reconstruction techniques in impulsive environments, while providing comparable performance in less demanding, light-tailed environments. Numerical results also demonstrate that the modifications improve LIHT performance, thereby requiring fewer samples to yield an approximate reconstruction.

The organization of the Chapter is as follows. Section 6.2 gives a brief review of CS and motivates the need for a simple robust algorithm capable of prior support knowledge inclusion. In Section 6.3 a robust iterative algorithm based on the Lorentzian norm is proposed and its properties are analyzed. In Section 6.4 we propose simple modifications for the developed algorithm to include prior signal information and analyze the partially known support case. Numerical experiments

evaluating the performance of the proposed algorithms in different environments are presented in Section 6.5. Finally, we close in Section 6.6 with conclusions and future directions.

6.2 Background and Motivation

6.2.1 Lorentzian Based Basis Pursuit

Let $x \in \mathbb{R}^n$ be an s -sparse signal or an s -compressible signal. A signal is s -sparse if only s of its coefficients are nonzero (usually $s \ll n$). A signal is s -compressible if its ordered set of coefficients decays rapidly and x is well approximated by the first s coefficients [47].

Let Φ be an $m \times n$ sensing matrix, $m < n$, with rows that form a set of vectors incoherent with the sparsity basis [47]. The signal x is measured by $y = \Phi x + z$, where z is the measurement (sampling) noise. It has been shown that a linear program (Basis Pursuit) can recover the original signal, x , from y [47]. However, there are several reconstruction methods that yield perfect or approximate reconstructions proposed in the literature (see [34,47,63,134] and references therein). Most CS algorithms use the L_2 norm as the metric for the residual error. However, it is well-known that LS based estimators are highly sensitive to outliers present in the measurement vector, leading to a poor performance when the noise no longer follows the Gaussian assumption but, instead, is better characterized by heavier-than-Gaussian tailed distributions [57,142].

In [57] we propose a robust reconstruction approach coined Lorentzian basis pursuit (BP). This method is a robust algorithm capable of reconstructing sparse signals in the presence of impulsive sampling noise. We use the following non-linear optimization problem to estimate x_0 from y :

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \text{ subject to } \|y - \Phi x\|_{LL_2, \gamma} \leq \epsilon \quad (6.1)$$

where

$$\|u\|_{LL_2, \gamma} = \sum_{i=1}^m \log\{1 + \gamma^{-2} u_i^2\}, \quad u \in \mathbb{R}^m, \quad \gamma > 0, \quad (6.2)$$

is the Lorentzian, or LL_2 , norm. The LL_2 norm does not over penalize large deviations, and is therefore a robust metric appropriate for impulsive environments [57]. The performance analysis of the algorithm is based on the so called restricted isometry properties (RIP) of the matrix Φ [47], which are defined in the following.

Definition 10. *The s -restricted isometry constant of Φ , δ_s , is defined as the smallest positive quantity such that*

$$(1 - \delta_s)\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1 + \delta_s)\|v\|_2^2$$

holds for all $v \in \Omega_s$, where $\Omega_s = \{v \in \mathbb{R}^n \mid \|v\|_0 \leq s\}$. A matrix Φ is said to satisfy the RIP of order s if $\delta_s \in (0, 1)$.

Carrillo *et. al* show in [57] that if Φ meets the RIP of order $2s$, with $\delta_{2s} < \sqrt{2} - 1$, then for any s -sparse signal x_0 and observation noise z with $\|z\|_{LL_2, \gamma} \leq \epsilon$, the solution to (6.1), denoted as x^* , obeys

$$\|x^* - x_0\|_2 \leq C_s \cdot 2\gamma \cdot \sqrt{m(e^\epsilon - 1)}, \quad (6.3)$$

where C_s is a small constant. One remark is that γ controls the robustness of the employed norm and ϵ the radius of the feasibility set LL_2 ball.

Although Lorentzian BP outperforms state of the art CS recovery algorithms in impulsive environments, while achieving comparable performance in less demanding light-tailed environments, numerical algorithms to solve the optimization problem posed by Lorentzian BP are extremely slow and complex [57]. Therefore, faster and simpler methods are sought to solve the sparse recovery problem in the presence of impulsive sampling noise.

6.2.2 Iterative hard thresholding

The iterative hard thresholding (IHT) algorithm is a simple iterative method that does not require matrix inversion at any point and provides near-optimal error guarantees [33, 34]. The algorithm is described as follows.

Let $x^{(t)}$ denote the solution at iteration time t and set $x^{(0)}$ to the zero vector. At each iteration t the algorithm computes

$$x^{(t+1)} = H_s(x^{(t)} + \mu\Phi^T(y - \Phi x^{(t)})), \quad (6.4)$$

where $H_s(a)$ is the non-linear operator that sets all but the largest (in magnitude) s elements of a to zero and μ is a step size. If there is no unique largest set, a set can be selected either randomly or based on a predefined ordering. Convergence of this algorithm is proven in [32] under the condition that $\|\Phi\|_{2 \rightarrow 2} < 1$, where $\|\Phi\|_{2 \rightarrow 2}$ represents the spectral norm of Φ , and a theoretical analysis for compressed sensing problems is presented in [33, 34]. Blumensath and Davies showed in [33] that if $\|z\|_2 \leq \epsilon$ (L_2 bounded noise) and $\delta_{3s} < 1/\sqrt{32}$, the reconstruction error of the IHT algorithm at iteration t is bounded by

$$\|x - x^{(t)}\|_2 \leq \alpha^t \|x\|_2 + \beta\epsilon, \quad (6.5)$$

where $\alpha < 1$ and β are absolute constants that depend only on δ_{2s} and δ_{3s} .

6.2.3 Compressed sensing with partially known support

Recent works show that modifying the CS framework to include prior knowledge of the support improves the reconstruction results using fewer measurements [119, 164]. Let $x \in \mathbb{R}^n$ be an sparse or compressible signal in some basis Ψ and denote $T = \text{supp}(x)$. In this setting, we assume that T is partially known, *i.e.* $T = T_0 \cup \Delta$. The set $T_0 \subset \{1, \dots, n\}$ is the *a priori* knowledge of the support

of x and $\Delta \subset \{1, \dots, n\}$ is the unknown part of the support. This scenario is typical in many real signal processing applications, *e.g.*, the lowest subband coefficients in a wavelet decomposition, which represent a low frequency approximation of the signal, or the first coefficients of a DCT transform of an image with a constant background, are known to be significant components.

The *a priori* information modified CS seeks out a signal that explains the measurements and whose support contains the smallest number of new additions to T_0 . Vaswani *et al.* proposed in [164] to modify BP to find an sparse signal assuming uncorrupted measurements. This technique is extended by Jacques in [119] to the case of corrupted measurements and compressible signals. Jacques finds sufficient conditions in terms of RIP for stable reconstruction in this general case. The approach solves the following optimization program

$$\min_{x \in \mathbb{R}^n} \|x_{T_0^c}\|_1 \quad \text{s. t.} \quad \|y - \Phi x\|_2 \leq \epsilon, \quad (6.6)$$

where x_Ω denotes the vector x with everything except the components indexed in $\Omega \subset \{1, \dots, n\}$ set to 0.

Although the modified CS approach needs fewer samples to recover a signal, the computational cost of solving (6.6) can be high, or complicated to implement. Therefore, in [58] we proposed to extend the ideas of modified CS to iterative approaches like greedy algorithms [134, 161] and iterative re-weighted least squares methods [53] (see Appendix C). Even though the aforementioned methods are more efficient than BP, in terms of computational cost, a disadvantage is that these methods need to invert a linear system at each iteration. In the following section we develop a robust algorithm inspired by the IHT algorithm that is capable of diminishing the effect of impulsive noise while able to including partial support information.

6.3 Lorentzian based Iterative Hard Thresholding Algorithm

In this section we propose a Lorentzian derived IHT algorithm for the recovery of sparse signals when the measurements are (possibly) corrupted by impulsive noise. First, we present the algorithm formulation and derive theoretical guarantees. Then, we describe how to optimize the algorithm parameters for enhanced performance.

6.3.1 Algorithm formulation and stability guarantees

Let $x_0 \in \mathbb{R}^n$ be an s -sparse or s -compressible signal, $s < n$. Consider again the sampling model

$$y = \Phi x_0 + z,$$

where Φ is an $m \times n$ sensing matrix and z denotes the sampling noise vector. In order to estimate x_0 from y we pose the following optimization problem:

$$\min_{x \in \mathbb{R}^n} \|y - \Phi x\|_{LL_2, \gamma} \quad \text{subject to} \quad \|x\|_0 \leq s. \quad (6.7)$$

However, the problem in (6.7) is non-convex and combinatorial. Therefore we derive a suboptimal strategy to estimate x_0 based on the gradient projection algorithm [28]. The proposed strategy is formulated as follows. Let $x^{(t)}$ denote again the solution at iteration time t and set $x^{(0)}$ to the zero vector. At each iteration t the algorithm computes

$$x^{(t+1)} = H_s (x^{(t)} + \mu g^{(t)}) \quad (6.8)$$

where

$$g = -\nabla_x \|y - \Phi x\|_{LL_2, \gamma}.$$

The negative gradient, g , can be expressed in the following form. Denote ϕ_i as the i -th row vector of Φ . Then

$$g^{(t)} = \Phi^T W_t (y - \Phi x^{(t)}) \quad (6.9)$$

where W_t is an $m \times m$ diagonal matrix with each element on the diagonal defined as

$$[W_t]_{i,i} = \frac{\gamma^2}{\gamma^2 + (y_i - \phi_i^T x^{(t)})^2}, \quad i = 1, \dots, m. \quad (6.10)$$

We coin the algorithm defined by the update in (6.8) as Lorentzian iterative hard thresholding (LIHT). The derived algorithm is almost identical to LS based IHT in terms of computational load, except for the additional cost of computing the m weights in (6.10) and a multiplication by an $m \times m$ diagonal matrix. For this additional cost we gain the advantage of robustness against heavy-tailed impulsive noise. Therefore the computational complexity per iteration of LIHT remains $\mathcal{O}(mn)$, which is limited by the matrix multiplication used. If fast matrix multiplication algorithms are available the complexity is reduced. Note that $[W_t]_{i,i} \leq 1$, with the weights going to zero when large deviations, compared to γ , are detected. In fact, if $W_t = I$ the algorithm reduces to the LS based IHT. Thus, the algorithm can be seen as a re-weighted least squares thresholding approach in which the weights diminish the effect of gross errors, assigning a small weight for large deviations and a weight near one for deviations close to zero. Fig. 6.1 shows an example of the obtained weight function with $\gamma = 1$.

In the following, we show that LIHT has theoretical stability guarantees similar to those of IHT. For simplicity of the analysis we set $\mu = 1$ as in [33].

Theorem 7. *Let $x_0 \in \mathbb{R}^n$. Define $S = \text{supp}(x_0)$, $|S| \leq s$. Suppose $\Phi \in \mathbb{R}^{m \times n}$ meets the RIP of order $3s$ and $\|\Phi\|_{2 \rightarrow 2} \leq 1$. Assume $x^{(0)} = 0$. Then if $\|z\|_{LL_2, \gamma} \leq \epsilon$ and $\delta_{3s} < 1/\sqrt{32}$ the reconstruction error of the LIHT algorithm at iteration t is bounded by*

$$\|x_0 - x^{(t)}\|_2 \leq \alpha^t \|x_0\|_2 + \beta \gamma \sqrt{m(e^\epsilon - 1)}, \quad (6.11)$$

where $\alpha = \sqrt{8}\delta_{3s}$ and $\beta = \sqrt{1 + \delta_{2s}}(1 - \alpha^t)(1 - \alpha)^{-1}$.

Proof of Theorem 7 follows from the fact that $W_t(i, i) \leq 1$, which implies

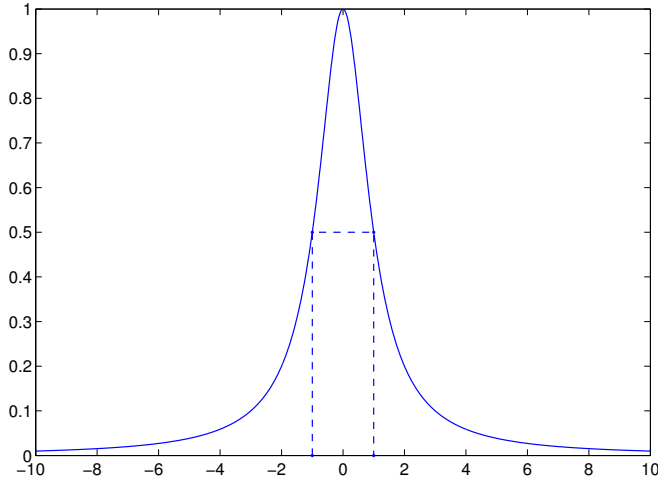


Figure 6.1: Weight function for $\gamma = 1$. Large deviations have a weight close to zero whilst small deviations have a weight close to one.

that

$$\|W_t z\|_2 \leq \|z\|_2 \leq \gamma \sqrt{m(e^\epsilon - 1)},$$

where the second inequality follows from Lemma 1 in [57]. Argument details parallel those of the proof of Theorem 8 in next section and, in fact, Theorem 7 is a particular case of Theorem 8. Therefore we provide only a proof for the later.

Although the algorithm is not guaranteed to converge to a global minima of (6.7), it can be shown that LIHT converges to a local minima since $[W_t]_{i,i} \leq 1$. Thus the eigenvalues of $\Phi^T W_t \Phi$ are bounded above by the eigenvalues of $\Phi^T \Phi$ and the sufficient condition $\|\Phi\|_{2 \rightarrow 2} \leq 1$ guarantees local convergence [33]. Notice that the RIP sufficient condition for stable recovery is identical to the one required by the LS based IHT algorithm [33].

The results in Theorem 7 can be easily extended to compressible signals using Lemma 6.1 in [134]. Suppose $x_0 \in \mathbb{R}^n$ is a s -compressible signal. Suppose $\Phi \in \mathbb{R}^{m \times n}$ meets the RIP of order $3s$ and $\|\Phi\|_{2 \rightarrow 2} \leq 1$. Assume $x^{(0)} = 0$. Then if the conditions of Theorem 7 are met, then the reconstruction error of the LIHT

algorithm at iteration t is bounded by

$$\|x_0 - x^{(t)}\|_2 \leq \eta \left(\|x_0 - x_s\|_2 + \frac{\|x_0 - x_s\|_1}{s} \right) + \alpha^t \|x_0\|_2 + \beta\gamma\sqrt{m(e^\epsilon - 1)}, \quad (6.12)$$

where $\alpha = \sqrt{8}\delta_{3s}$, $\beta = \sqrt{1 + \delta_{2s}}(1 - \alpha^t)(1 - \alpha)^{-1}$, $\eta = \sqrt{1 + \delta_s}$ and x_s is the best s -term approximation of x_0 .

6.3.2 Parameter tuning

The performance of the LIHT algorithm depends on the scale parameter γ of the Lorentzian norm and the step size, μ . Therefore, we detail methods to estimate these two parameters in the following.

It is observed in [57] that setting γ as half the sample range of y , $(y_{(1)} - y_{(0)})/2$ (where $y_{(q)}$ denotes the q -th quantile of y), often makes the Lorentzian norm a fair approximation to the L_2 norm. Therefore, the optimal value of γ should be $(y'_{(1)} - y'_{(0)})/2$, where $y' = \Phi x_0$ is the uncorrupted measurement vector. Since the uncorrupted measurements are unknown, we propose to estimate the scale parameter as

$$\gamma = \frac{y_{(0.875)} - y_{(0.125)}}{2}. \quad (6.13)$$

This value of γ implicitly considers a measurement vector with 25% of the samples corrupted by outliers and 75% well behaved. Experimental results show that this estimate leads to good performance in both Gaussian and impulsive environments (see Section 6.5).

As described in [34], the convergence and performance of the LS based IHT algorithm improves if an adaptive step size, $\mu^{(t)}$, is used to normalize the gradient update. We use a similar approach here. Let $S^{(t)}$ be the support of $x^{(t)}$ and suppose that the algorithm has identified the true support of x_0 , *i.e.* $S^{(t)} = S^{(t+1)} = S$. In this case we want to minimize $\|y - \Phi_S x_S\|_{LL_2, \gamma}$ using a gradient descent algorithm

with updates of the form

$$x_S^{(t+1)} = x_S^{(t)} + \mu^{(t)} g_S^{(t)}. \quad (6.14)$$

Finding the optimal μ , *i.e.*, a step size that maximally reduces the objective at each iteration is not an easy task and, in fact, there is no known closed form for such optimal step. To overcome this limitation we propose to use a suboptimal approach that still guarantees a reduction in the objective function in each iteration. We set the step size in each iteration as

$$\begin{aligned} \mu^{(t)} &= \min_{\mu} \left\| W_t^{1/2} \left[y - \Phi_S \left(x_S^{(t)} + \mu g_S^{(t)} \right) \right] \right\|_2^2 \\ &= \frac{\|g_S^{(t)}\|_2^2}{\|W_t^{1/2} \Phi_S g_S^{(t)}\|_2^2}, \end{aligned} \quad (6.15)$$

which guarantees that the objective Lorentzian function is not increased at each iteration.

Proposition 4. *Let $\mu^{(t)} = \|g_S^{(t)}\|_2^2 / \|W_t^{1/2} \Phi_S g_S^{(t)}\|_2^2$ and $x_S^{(t+1)} = x_S^{(t)} + \mu^{(t)} g_S^{(t)}$. Then, if $S^{(t)} = S^{(t+1)} = S$ the update guarantees that*

$$\|y - \Phi x^{(t+1)}\|_{LL_2, \gamma} \leq \|y - \Phi x^{(t)}\|_{LL_2, \gamma}.$$

Before proving Proposition 4, we need a known result for square concave functions that is used in the proof.

Proposition 5. *Let $f(a) = g(a^2)$ with g concave. Then for any $a, b \in \mathbb{R}$ we have the following inequality:*

$$f(a) - f(b) \leq \frac{f'(b)}{2b} (a^2 - b^2)$$

which is the differential criterion for the concavity of g .

Now we can proof Proposition 4.

Proof. Define

$$f(a) = \log \left(1 + \frac{a^2}{\gamma^2} \right) \quad \text{and} \quad r^{(t)} = y - \Phi x^{(t)}.$$

Using Proposition 5 and the fact that $f(x)$ is square concave, we have the following inequality:

$$\begin{aligned} \sum_{i=1}^m f([r^{(t+1)}]_i) - f([r^{(t)}]_i) &\leq \frac{1}{2} \sum_{i=1}^m \frac{f'([r^{(t)}]_i)}{[r^{(t)}]_i} ([r^{(t+1)}]_i^2 - [r^{(t)}]_i^2) \\ &= \frac{1}{2\gamma^2} \sum_{i=1}^m [W_t]_{ii} [r^{(t+1)}]_i^2 + \frac{1}{2\gamma^2} \sum_{i=1}^m [W_t]_{ii} [r^{(t)}]_i^2. \end{aligned}$$

This is equivalent to

$$\begin{aligned} \|y - \Phi x^{(t+1)}\|_{LL_2, \gamma} - \|y - \Phi x^{(t)}\|_{LL_2, \gamma} \\ \leq \frac{1}{2\gamma^2} \|W_t^{1/2}(y - \Phi x^{(t+1)})\|_2^2 - \frac{1}{2\gamma^2} \|W_t^{1/2}(y - \Phi x^{(t)})\|_2^2. \end{aligned}$$

From the optimality of $\mu^{(t)}$ we have

$$\|W_t^{1/2}(y - \Phi x^{(t+1)})\|_2^2 - \|W_t^{1/2}(y - \Phi x^{(t)})\|_2^2 \leq 0.$$

Therefore

$$\|y - \Phi x^{(t+1)}\|_{LL_2, \gamma} - \|y - \Phi x^{(t)}\|_{LL_2, \gamma} \leq 0$$

which is the desired result. \square

In the case in which the support of $x^{(t+1)}$ differs from the support of $x^{(t)}$, the optimality of $\mu^{(t)}$ is no longer guaranteed. If

$$\|y - \Phi x^{(t+1)}\|_{LL_2, \gamma} > \|y - \Phi x^{(t)}\|_{LL_2, \gamma},$$

we use a backtracking algorithm and set $\mu^{(t)} \leftarrow \mu^{(t)}/2$ until the objective function

in (6.7) is reduced.

6.4 Lorentzian Iterative Hard Thresholding with Prior Information

In this section we modify the LIHT algorithm to incorporate prior signal information into the recovery process. The LIHT algorithm constructs an estimate of the signal at each iteration, thereby enabling intuitive incorporation of prior knowledge in each step of the recursion. In the following we propose extensions of the LIHT algorithm to incorporate partial support knowledge and then describe a general modification to include the model-based CS framework of [22].

6.4.1 Lorentzian iterative hard thresholding with partially known support

Let $x_0 \in \mathbb{R}^n$ be an s -sparse or s -compressible signal, $s < n$. Consider the sampling model $y = \Phi x_0 + z$, where Φ is an $m \times n$ sensing matrix and z denotes the sampling noise vector. Denote $T = \text{supp}(x_0)$ and assume that T is partially known, *i.e.* $T = T_0 \cup \Delta$. Define $k = |T_0|$. We propose a simple extension of the LIHT algorithm that incorporates the partial support knowledge into the recovery process. The modification of the algorithm is described in the following.

Denote $x^{(t)}$ as the solution at iteration t and set $x^{(0)}$ to the zero vector. At each iteration t the algorithm computes

$$x^{(t+1)} = H_{s-k}^{T_0} \left(x^{(t)} + \mu^{(t)} \Phi^T W_t (y - \Phi x^{(t)}) \right), \quad (6.16)$$

where the nonlinear operator $H_u^\Omega(\cdot)$ is defined as

$$H_u^\Omega(a) = a_\Omega + H_u(a_{\Omega^c}), \quad \Omega \subset \{1, \dots, n\}. \quad (6.17)$$

The algorithm selects the $s-k$ largest (in magnitude) components that are not in T_0 and preserves all components in T_0 at each iteration. We coined this algorithm

Lorentzian iterative hard thresholding with partially known support (LIHT-PKS).

The main result of this section, Theorem 8 below, shows the stability of LIHT-PKS and establish sufficient conditions for stable recovery in terms of the RIP of Φ . In the following we show that LIHT-PKS has theoretical stability guarantees similar to those of IHT [33]. For simplicity of the analysis we set $\mu = 1$ as in section 6.3.

Theorem 8. *Let $x \in \mathbb{R}^n$. Define $T = \text{supp}(x)$ with $|T| = s$. Also define $T = T_0 \cup \Delta$ and $|T_0| = k$. Suppose $\Phi \in \mathbb{R}^{m \times n}$ meets the RIP of order $3s - 2k$ and $\|\Phi\|_{2 \rightarrow 2} \leq 1$. Then if $\|z\|_{LL_2, \gamma} \leq \epsilon$ and $\delta_{3s-2k} < 1/\sqrt{32}$, the reconstruction error of the IHT-PKS algorithm at iteration t is bounded by*

$$\|x_0 - x^{(t)}\|_2 \leq \alpha^t \|x\|_2 + \beta \gamma \sqrt{m(e^\epsilon - 1)}, \quad (6.18)$$

where

$$\alpha = \sqrt{8} \delta_{3s-2k} \quad \text{and} \quad \beta = \sqrt{1 + \delta_{2s-k}} \left(\frac{1 - \alpha^t}{1 - \alpha} \right).$$

Proof. See appendix D. □

A sufficient condition for stable recovery of the LIHT algorithm is $\delta_{3s} < 1/\sqrt{32}$ (see section 6.3), which is a stronger condition than that required by LIHT-PKS since $\delta_{3s-2k} < \delta_{3s}$. Having a RIP of smaller order means that Φ requires fewer rows to meet the condition, *i.e.*, fewer samples to achieve approximate reconstruction. Notice that when $k = 0$ (cardinality of the partially known support) we have the same condition required by LIHT. The results in Theorem 8 can be easily extended to compressible signals using Lemma 6.1 in [134], as was done in the previous section for LIHT.

6.4.2 Extension of Lorentzian iterative hard thresholding to model-sparse signals

Baraniuk *et. al* introduced a model-based CS theory that reduces the degrees of freedom of a sparse/compressible signal [22, 90]. The key ingredient of this approach is to use a more realistic signal model that goes beyond simple sparsity by codifying the inter-dependency structure among the signal coefficients. This signal model might be a wavelet tree, block sparsity or in general a union of s -dimensional subspaces [22].

Suppose \mathcal{M}_s is a signal model as defined in [22] and also suppose that $x_0 \in \mathcal{M}_s$ is an s -model sparse signal. Then, a model-based extension of the LIHT algorithm is motivated by solving the problem

$$\min_{x \in \mathcal{M}_s} \|y - \Phi x\|_{LL_2, \gamma}, \quad (6.19)$$

using the following recursion:

$$x^{(t+1)} = \mathbb{M}_s \left(x^{(t)} + \mu^{(t)} \Phi^T W_t (y - \Phi x^{(t)}) \right), \quad (6.20)$$

where $\mathbb{M}_s(a)$ is the best s -term model-based operator that projects the vector a onto \mathcal{M}_s . One remark to make is that under the model-based CS framework of [22] this prior knowledge model can be leveraged in recovery with the resulting algorithm being similar to LIHT-PKS.

6.5 Experimental Results

6.5.1 Robust Reconstruction: LIHT

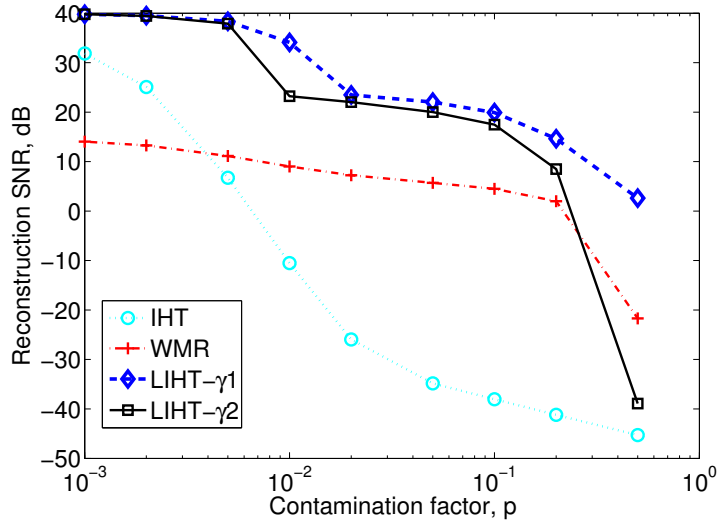
Numerical experiments that illustrate the effectiveness of the LIHT algorithm are presented in this section. All experiments utilize synthetic s -sparse signals in a Hadamard basis, with $s = 8$ and $n = 1024$. The nonzero coefficients have equal

amplitude, equiprobable sign, randomly chosen position, and average power fixed to 0.78. Gaussian sensing matrices are employed with $m = 128$. One thousand repetitions of each experiment are averaged and reconstruction SNR is used as the performance measure. Weighted median regression (WMR) [142] and LS-IHT [34] are used as benchmarks.

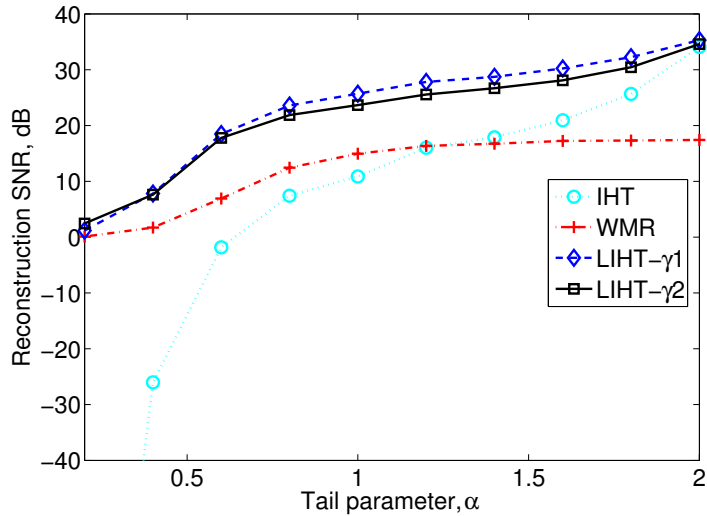
To test the robustness of the methods, we use two noise models: α -stable distributed noise and Gaussian noise plus gross sparse errors. The Gaussian noise plus gross sparse errors model is referred to as contaminated p -Gaussian noise for the remainder of the paper, as p represents the amount of gross error contamination. To validate the estimate of γ discussed in Section 6.3.2 we make a comparison between the performance of LIHT equipped with the optimal γ , denoted as LIHT- γ_1 , and the signal-estimated γ , denoted as LIHT- γ_2 . The optimal γ is set as half the sample range of the clean measurements.

For the first experiment we consider a mixed noise environment, using contaminated p -Gaussian noise. We set the Gaussian component variance to $\sigma^2 = 10^{-2}$, resulting in an SNR of 18.9321 dB when $p = 0$. The amplitude of the outliers is set as $\delta = 10^3$ and p is varied from 10^{-3} to 0.5. The results are shown in Fig. 6.2 (a). The results demonstrate that LIHT outperforms WMR and IHT. Moreover, the results also demonstrate the validity of the estimated γ . Although the reconstruction quality achieved by LIHT- γ_2 is lower than that achieved by LIHT- γ_1 , the SNR of LIHT- γ_2 is greater than 20 dB for a broad range of contamination factors p , including contaminations up to 5% of the measurements.

The second experiment explores the behaviour of LIHT in very impulsive environments. We compare again against IHT and WMR, this time with α -Stable sampling noise. The scale parameter of the noise is set as $\sigma = 0.1$ for all cases and the tail parameter, α , is varied from 0.2 to 2, *i.e.*, very impulsive to the Gaussian case, Fig. 6.2 (b). For small values of α , all methods perform poorly, with LIHT



(a)



(b)

Figure 6.2: Comparison of LIHT with LS-IHT and WMR for impulsive contaminated samples. (a) Contaminated p -Gaussian, $\sigma^2 = 0.01$. R-SNR as a function of the contamination parameter, p . (b) α -stable noise, $\sigma = 0.1$. R-SNR as a function of the tail parameter, α .

yielding the most acceptable results. Beyond $\alpha = 0.6$, LIHT produces faithful reconstructions with a SNR greater than 20 dB, and often 10 dB greater than IHT and WMR results. It is of note that when $\alpha = 2$ (Gaussian case) the performance of LIHT is comparable with that of IHT, which is least squares based. Also of notice is that the SNRs achieved by LIHT- γ_1 and LIHT- γ_2 are almost identical, being LIHT- γ_1 slightly better.

For the next experiment, we evaluate the performance of LIHT as the number of measurements varies for different levels of impulsiveness. The number of measurements is varied from 16 (twice the sparsity level) to 512 (half the dimension of x_0). The sampling noise model used is α -stable with four values of α : 0.5, 1, 1.5, 2. The results are summarized in Fig. 6.3, which show that, for $\alpha \in [1, 2]$, LIHT yields fair reconstructions from 96 samples. However for $\alpha = 0.5$ (most impulsive case of the four), more samples are needed, 256, to yield a fair reconstruction. Results of IHT with Gaussian noise ($\alpha = 2$) are also included for comparison. It is of note that the performance of LIHT is comparable to that of IHT for the Gaussian case.

The last experiment in this subsection shows the effectiveness of LIHT to recover real signals from corrupted measurements. We take random Hadamard measurements of the the 256×256 ($n = 65536$) Lena image and then we add Cauchy distributed noise to the measurements. We fix the number of measurements as $m = 32000$ and the scale (dispersion) parameter of the Cauchy noise to $\sigma = 1$. Fig. 6.4 shows the clean measurements on the top image and the Cauchy corrupted measurements in the bottom one.

We compare the reconstruction results of LIHT to those obtained by the classical least squares IHT (LS-IHT) algorithm and the LS-IHT with noise clipping, which is the classical approach to reject outliers. To set a clipping rule we assume that we know before hand the the range of the clean measurements and all samples

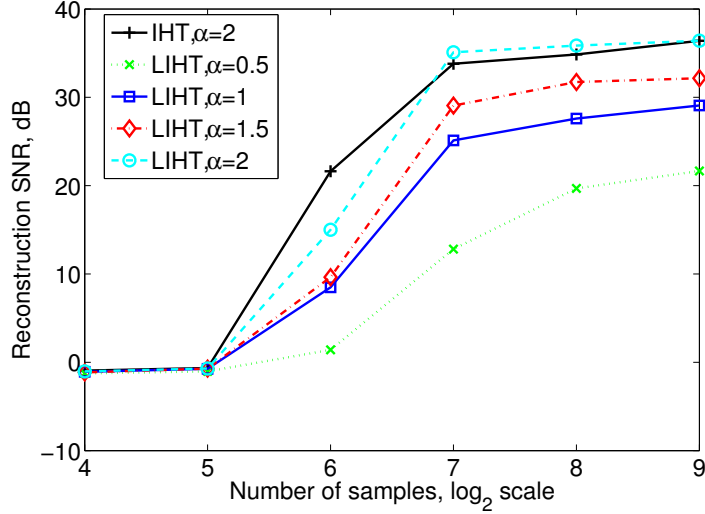


Figure 6.3: Performance of LIHT as the number of measurements varies for synthetic sparse signals. Reconstruction SNR as a function of the number of measurements.

are clipped within this range, *i.e.*

$$y_i^c = \begin{cases} y_{min}, & \text{if } y_i \leq y_{min} \\ y_i, & \text{if } y_{min} < y_i < y_{max} \\ y_{max}, & \text{if } y_i \geq y_{max} \end{cases}$$

where y^c denotes the vector of clipped measurements. For LIHT we estimate γ using equation (6.13). For all experiments we assume a sparsity level of $s = 6000$. Fig. 6.5 (a) shows the reconstructed image using LS-IHT, R-SNR=-10.7 dB, Fig. 6.5 (b) shows the reconstructed image using LS-IHT and noise clipping, R-SNR=6.2 dB and Fig. 6.5 (c) shows the reconstructed image using LIHT, R-SNR=20.5 dB. Fig. 6.5 (d) shows the reconstructed image from noiseless measurements using LS-IHT as comparison, R-SNR=23.9 dB. From the results is clear that LIHT outperform the other approaches and the reconstruction quality is about 3 dB worse than the noiseless reconstruction. Furthermore, the results of LS-IHT with a clipping

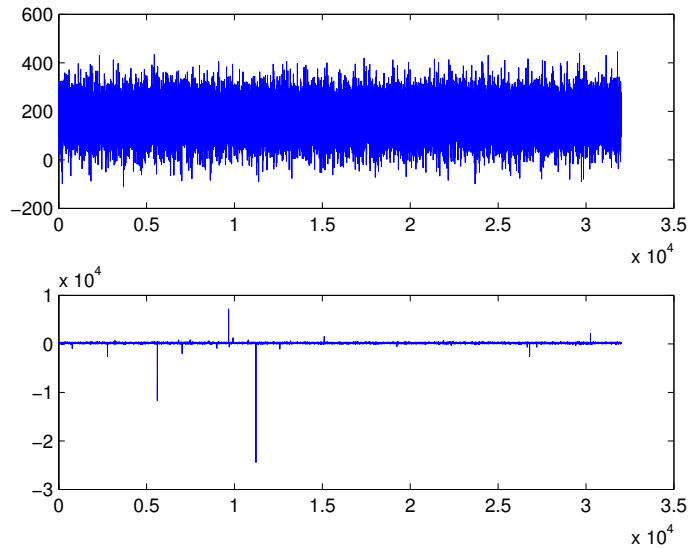


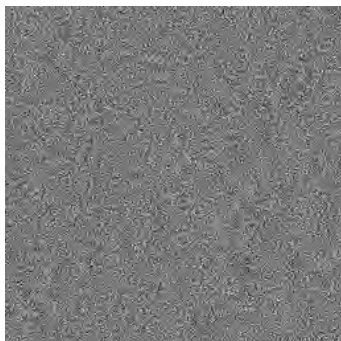
Figure 6.4: Example of a 256×256 image sampled by a random Hadamard ensemble. Top: clean measurements. Bottom: Cauchy corrupted measurements, $\sigma = 1$.

strategy, even with the clean measurements range as prior information, are not as expected showing the superiority of robust operators in impulsive environments.

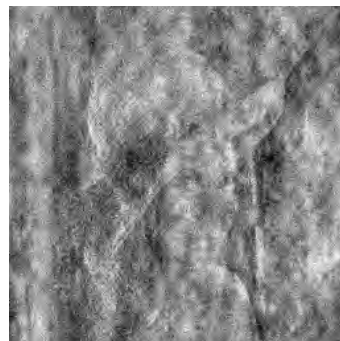
6.5.2 LIHT with Partially Known Support

Numerical experiments that illustrate the effectiveness of LIHT with partially known support are presented in this section. Results are presented for synthetic and real signals. In the real signal case, comparisons are made with a broad set of alternative algorithms.

Synthetic sparse vectors are employed in the first experiment. The signal length is set as $n = 1000$ and the sparsity level is fixed to 50. The nonzero coefficients are drawn from a Rademacher distribution, their position randomly chosen and amplitudes $\{-10, 10\}$. The vectors are sampled using sensing matrices Φ that have i.i.d. entries drawn from a standard normal distribution with normalized columns. Each experiment is repeated 300 times, with average results presented.



(a)



(b)



(c)



(d)

Figure 6.5: Lena image reconstruction example from measurements corrupted by Cauchy noise. (a) Reconstructed image using LS-IHT, R-SNR=-10.7 dB. (b) Reconstructed image using LS-IHT and noise clipping, R-SNR=6.2 dB. (c) Reconstructed image using LIHT, R-SNR=20.5 dB. (d) Reconstructed image from noiseless measurements using LS-IHT, R-SNR=23.9 dB.

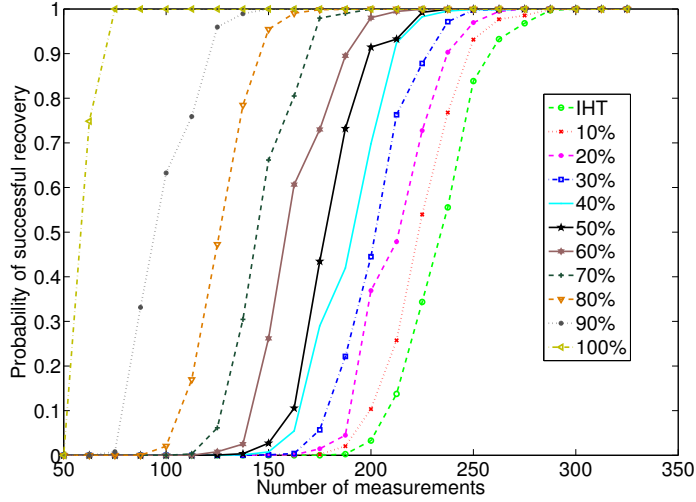


Figure 6.6: Probability of successful recovery as a function of the number of measurements, for different percentages of partially known support.

The effect of including partial support knowledge is analyzed by increasing the cardinality of the known set in steps of 10% for different numbers of measurements. The probability of exact reconstruction is employed as a measure of performance. Fig. 6.6 shows that, as expected, the reconstruction accuracy grows with the percentage of known support. The results also show that incorporating prior support information substantially reduces the number of measurements required for successful recovery.

The second experiment illustrates algorithm performance for real compressible signals. ECG signals are utilized due to the structure of their sparse decompositions. Experiments are carried out over 10-min long leads extracted from records 100, 101, 102, 103, 107, 109, 111, 115, 117, 118 and 119 from the MIT-BIH Arrhythmia Database (see [30] and references therein). Cosine modulated filter banks are used to determine a sparse representation of the signal [30]. A sparse signal approximation is determined by processing 1024 samples of ECG data, setting the number of channels, M , to 16, and selecting the largest 128 coefficients. This support set is

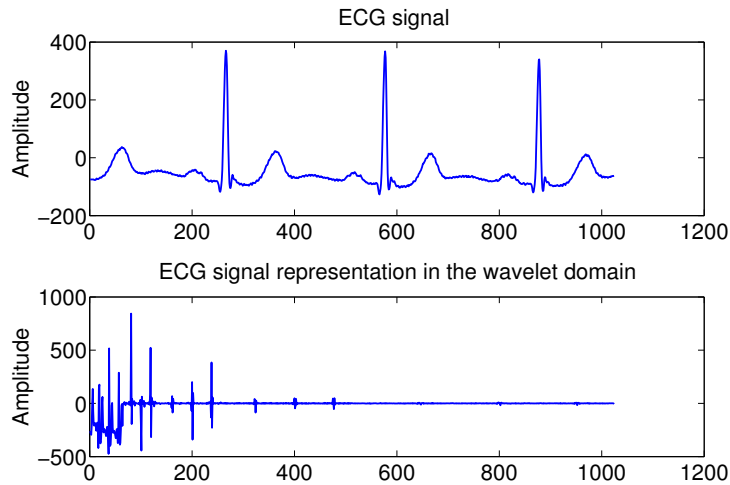


Figure 6.7: Decomposition of an ECG signal using CMFB, $M = 16$ and $n = 1024$.

denoted by T ; note that $|T| = 128$. Fig. 6.7 shows an example of a decomposition of a lead of 1024 samples and its decomposition using CMFB.

Three cases are considered. In the first, the median (magnitude) support coefficient is determined and the coefficients of T with magnitudes greater than or equal to the median are designated as the known signal support, *i.e.*, the positions of the largest (magnitude) 50% of T coefficients are taken to be the known signal support. This case is denoted as IHT-PKS-I. The second partially known support case corresponds to those with magnitude less than the median, *i.e.*, the positions of the smallest (magnitude) 50% of T coefficients since these might be the most difficult to find coefficients. This case is denoted as IHT-PKS-II. The third and final selection, denoted as IHT-PKS, is related to the low-pass approximation of the first subband, which corresponds to the first 64 coefficients (when $n = 1024$). This first subband accumulates the majority of signal energy, which is the motivation for this case.

Fig. 6.8 compares the three proposed partially known support selections. Each method improves the performance over standard LIHT, except for IHT-PKS-II

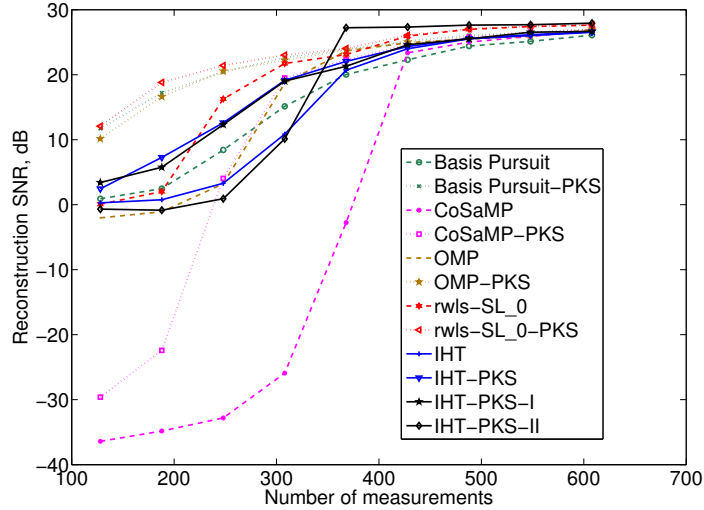


Figure 6.8: Comparison of LIHT, BP, OMP, CoSaMP, rwls-SL_0 and their partially known support versions for ECG signals.

when the number of measurements is not sufficient to achieve accurate reconstruction. Note, however, that the performance of IHT-PKS-II improves rapidly as the number of measurements increases, with the method outperforming the other algorithms in this regime. The performance of IHT-PKS-I is very similar to IHT-PKS since most of the first subband low-pass approximation coefficients are included in the 50% largest coefficients of T set. Notice that IHT-PKS-I performs slightly better than IHT-PKS for small numbers of measurements.

Also compared with LIHT in Fig. 6.8 are the OMP, CoSaMP, and rwls-SL_0 iterative algorithms, as well as their partially known support versions (OMP-PKS, CoSaMP-PKS, and rwls-SL_0 -PKS) [58]. For reference, we also include Basis Pursuit (BP) and Basis Pursuit with partially known support (BP-PKS) [164]. In all cases, the positions of the first subband low-pass approximation coefficients are selected as the signal partially known support. Note that LIHT-PKS performs better than CoSaMP-PKS for small numbers of measurements and yields similar reconstructions when the number of measurements increases. Although the known support versions

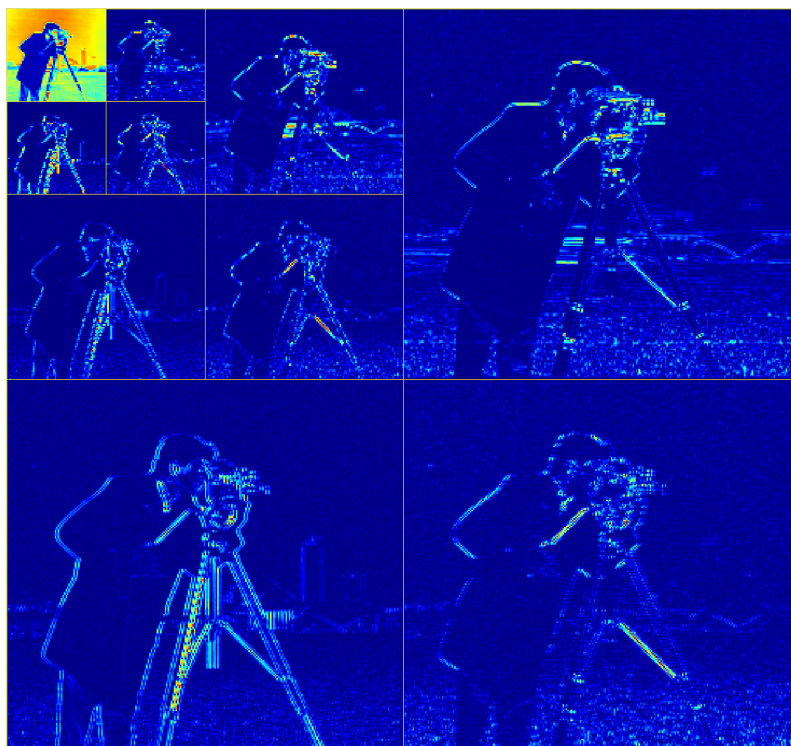


Figure 6.9: Wavelet decomposition of the camera man image.

of the other iterative algorithms require fewer measurements to achieve accurate reconstructions, LIHT does not require the exact solution to an inverse problem, thus making it computationally more efficient. And as in the previous example, the performance of Lorentzian iterative hard thresholding is improved through the inclusion of partially known support information, thereby enabling the number of measurements requires for a specified level of performance to be reduced.

As a final example we illustrate how the partially known support framework can be applied in image reconstruction. Consider the wavelet decomposition of the 256×256 camera man image shown in Fig. 6.9. We use Daubechies DB8 wavelets as our sparsity basis. The wavelet decomposition shows that for natural images the largest coefficients are concentrated in the approximation band and the remainder signal, detail coefficients, is a sparser signal than the original decomposition. Thus,

a possible form to incorporate the partially known support framework is to assume that the approximation band coefficients are part of the true signal support, *i.e.*, the partially known support.

To test our assumption we take random Hadamard measurements of the the 256×256 Lena image and then we estimate the image from the measurements. Fig. 6.10 top left shows the original image. Again we use Daubechies DB8 wavelets as our sparsity basis and we approximate the image with the largest 6000 coefficients, *i.e.*, $|T| = 6000$. Fig. 6.10 top right shows the best s -term approximation, $s = 6000$ with R-SNR=23.9 dB. We take $m = 16000$ measurements and reconstruct the image using the LIHT and LIHT-PKS algorithm. For LIHT-PKS we assume that the approximation band is in the true support of the image coefficients, $k = 2048$ for this example. The reconstruction results are shown in Fig. 6.10 bottom left and Fig. 6.10 bottom right respectively. The reconstruction SNR results are R-SNR=10.2 dB for the standard LIHT and R-SNR=20.4 dB for LIHT-PKS. The LIHT-PKS algorithm outperforms its counterpart without support knowledge by 10 dB. More importantly, the partially known support reconstruction quality is 3 dB below the reconstruction quality obtained by the best s -term approximation.

6.6 Concluding Remarks

This Chapter presents a Lorentzian based IHT algorithm for recovery of sparse signals in impulsive environments. The derived algorithm is comparable to least squares based IHT in terms of computational load with the advantage of robustness against heavy-tailed impulsive noise. Sufficient conditions for stability are studied and a reconstruction error bound is derived that depends on the noise strength and a tunable parameter of the Lorentzian norm. Methods to estimate the adjustable parameters of the reconstruction algorithm are also proposed. Simulations results show that the Lorentzian-based IHT algorithm yields comparable



Figure 6.10: Top left: Original 256×256 image. Top right: Best s -term approximation, $s = 6000$, R-SNR=23.9 dB. Reconstruction from $m = 16000$. Bottom left: LIHT, R-SNR=10.2 dB. Bottom right: LIHT-PKS $k = 2000$, R-SNR=20.4 dB.

performance with state of the art algorithms in light-tailed environments while having substantial performance improvements in heavy-tailed environments.

Additionally, this Chapter proposes a modification of the Lorentzian iterative hard thresholding algorithm that incorporates known support in the recovery process. Sufficient conditions for stable recovery in the compressed sensing with partially known support problem are derived. The theoretical analysis shows that including prior support information relaxes the conditions for successful reconstruction. Numerical results show that the LIHT modification improves performance, thereby requiring fewer samples to yield an approximate reconstruction.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

This dissertation investigates robust sensing and reconstruction methods for sparse signals in the compressed sensing (CS) framework. To achieve this goal, we make use of robust statistics theory to develop appropriate methods addressing the problem of impulsive noise in CS systems. The work in this dissertation can have significant impact in problems where the processes are corrupted by outliers, *e.g.*, missing or saturated samples. Examples of such problems are: channel coding for erasure channels, real image and data acquisition systems, atmospheric and underwater communications, computer networks, bioinformatics, medical imaging and geosciences. The contributions of this thesis are concentrated in three areas.

- *Robust signal processing*: robust estimation and filtering methods, as well as robust error metrics are developed from the GCD family.
- *Compressive sensing methods in impulsive noise*: robust sampling operators, together with robust reconstruction strategies are developed and their properties analyzed.
- *Compressive sensing with prior information*: fast reconstruction strategies that incorporate probabilistic signal models as well as deterministic signal prior information into the recovery process are developed.

Chapter 3 presents a GCD based theoretical approach that allows the formulation of challenging problems in a robust fashion. Within this framework, we establish a statistical relationship between the GGD and GCD families in Lemma 1. The proposed framework, due to its flexibility, subsumes GGD based developments, thereby guaranteeing performance improvements over the traditional problem formulation techniques. The developed theoretical framework includes robust estimation and filtering methods, as well as robust error metrics. Robust metric functions have a great impact in sparse reconstruction techniques, both as error metrics and as sparsity encouraging techniques [53, 103, 140, 167]. Properties of the derived techniques are analyzed. Three particular applications are developed under this framework: 1) robust filtering for power line communications, 2) robust estimation in sensor networks with noisy channels and 3) robust fuzzy clustering. Results from the applications show that the proposed GCD-derived methods provide a robust framework in impulsive heavy-tailed environments, with performance comparable to existing methods in less demanding light-tailed environments.

Chapter 4 presents robust sampling and reconstruction methods for sparse signals in impulsive environments. Myriad projections are proposed as sampling operators to address problems with impulsive observation noise. Properties of the proposed sampling function are analyzed, and it is noted that reconstruction performance depends on a linearity parameter, K , which can be adapted to the signal and noise environment. Importantly, myriad projections can be used with standard Gaussian-derived reconstruction algorithms. To address the problem of heavy-tailed sampling noise, Lorentzian basis pursuit is proposed. A reconstruction bound is derived that depends on the noise strength and a tunable parameter of the Lorentzian norm. Methods to estimate the adjustable parameters in the sampling functions and reconstruction algorithms are proposed, although computation of their optimal values remains an open question. Thus Myriad projections and Lorentzian

BP offer a robust framework for CS in impulsive heavy-tailed environments, with performance comparable to existing methods in less demanding light-tailed environments. Although the method outperforms state of the art CS recovery algorithms in impulsive environments and achieves comparable performance in less demanding light-tailed environments, numerical algorithms to solve the optimization problem posed by Lorentzian BP are extremely slow and complex. Therefore, faster and simpler methods are sought to solve the sparse recovery problem in the presence of impulsive sampling noise.

Chapter 5 formulates the CS recovery problem in a Bayesian framework using algebraic-tailed priors from the GCD family for the signal coefficients and the measurement noise. We show that algebraic-tailed impulsive distributions are more suitable models for sparse or compressible signals a conclusion also shown in [111]. An iterative reconstruction algorithm, referred to as GCBCS, is developed from this Bayesian formulation. Simulation results show that the proposed method requires fewer samples than most existing reconstruction algorithms for compressed sensing, thereby validating the use of GCD priors for sparse reconstruction problems. The proposed Bayesian yields comparable performance with state of the art algorithms in light-tailed noise environments while having substantial performance improvements in heavy-tailed environments.

Chapter 6 presents a Lorentzian based IHT algorithm for recovery of sparse signals in impulsive environments. The derived algorithm is comparable to least squares based IHT in terms of computational load with the advantage of robustness against heavy-tailed impulsive noise. Sufficient conditions for stability are studied and a reconstruction error bound is derived that depends on the noise strength and a tunable parameter of the Lorentzian norm. Simulations results show that the Lorentzian-based IHT algorithm yields comparable performance with state of the art algorithms in light-tailed environments while having substantial performance

improvements in heavy-tailed environments. Methods to estimate the adjustable parameters in the reconstruction algorithm are proposed, although computation of their optimal values remains an open question. Future work will focus on convergence analysis of the proposed algorithm. Additionally, Chapter 6 proposes a modification of the Lorentzian iterative hard thresholding algorithm that incorporates partially known support in the recovery process. Sufficient conditions for stable recovery in the compressed sensing with partially known support problem are derived. The theoretical analysis shows that including prior support information relaxes the conditions for successful reconstruction. Numerical results show that the LIHT modification improves performance, thereby requiring fewer samples to yield an approximate reconstruction. We also make a general formulation of the LIHT algorithm using the model-based CS framework of [22].

7.2 Future Work

There are many roads to follow for future work on the topics of this dissertation.

While myriad projections propose a robust framework for sampling signals in the presence of impulsive noise, its implementation is not natural and requires previous sampling of the input signal. Therefore, more natural nonlinear sampling operators (sensing procedures) should be investigated. One step in this direction is the work of Blumensath in [31], where he introduces a further generalization to compressed sensing and allow for non-linear sampling methods. As opposed to the work developed in this dissertation, where we try to approximate in the limit the nonlinear measurements by linear measurements, this work opens new roads for general nonlinear sampling systems. This generalization is achieved by using a recently introduced generalization of the Restricted Isometry Property (or the bi-Lipschitz condition) traditionally imposed on the compressed sensing system. The

author shows that, if this more general condition holds for the nonlinear sampling system, then we can reconstruct signals from non-linear compressive measurements.

Algebraic-tailed priors have received a lot of attention recently due to the fact that they pose concave optimization problems and numerical results show that these concave problems yield better signal estimates with the same number of measurements [46, 50, 53, 63, 73, 111, 152]. However, with the exception of [63], little work has been done in understanding this phenomena and a theoretical analysis is needed to show why the number of measurements is reduced. Therefore a theoretical analysis, either based on RIP or from a Bayesian perspective is needed. Also, most models considered in the literature for algebraic priors assume an i.i.d. structure of the coefficients, thereby not exploiting the intra signal correlation structure. One problem with the correlation approach is that algebraic distributions have infinite second moment, thus there is no straight application of the correlation concept. However alternative strategies can be developed to describe the coefficient structure and therefore achieve a lower sampling rate.

BIBLIOGRAPHY

- [1] Compressive sensing resources. Online: <http://www.dsp.ece.rice.edu/cs/>.
- [2] M. Akcakaya and V. Tarokh. A frame construction and a universal distortion bound for sparse representations. *IEEE Transactions on Signal Processing*, 56(6):2443–2550, June 2008.
- [3] G. R. Arce. A general weighted median filter structure admitting negative weights. *IEEE Transactions on Signal Processing*, 46:3195–3205, December 1998.
- [4] G. R. Arce. *Nonlinear Signal Processing: A Statistical Approach*. John Wiley & Sons, Inc., 2005.
- [5] G. R. Arce, D. Otero, A. B. Ramirez, and J. Paredes. Reconstruction of sparse signals from l_1 dimensionality-reduced cauchy random-projections. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, March 2010.
- [6] J. Astola and Y. Neuvo. Optimal median type filters for exponential noise distributions. *Signal Processing*, 17(2):95 – 104, June 1989.
- [7] T. C. Aysal. *Filtering and Estimation Theory: First-Order, Polynomial and Decentralized Signal Processing*. Ph.D. dissertation, ECE Department, University of Delaware, 2007.
- [8] T. C. Aysal and K. E. Barner. Hybrid polynomial filters for Gaussian and non-Gaussian noise environments. *IEEE Transactions on Signal Processing*, 54(12):4644–4661, December 2006.
- [9] T. C. Aysal and K. E. Barner. Meridian filtering for robust signal processing. *IEEE Transactions on Signal Processing*, 55(8):3949–3962, August 2007.
- [10] T. C. Aysal and K. E. Barner. Myriad-type polynomial filtering. *IEEE Transactions on Signal Processing*, 55(12):747–753, February 2007.

- [11] T. C. Aysal and K. E. Barner. Blind decentralized estimation for bandwidth constrained wireless sensor networks. *IEEE Transactions on Wireless Communications*, 7(5):1466–1471, May 2008.
- [12] T. C. Aysal and K. E. Barner. Constrained decentralized estimation over noisy channels for sensor networks. *IEEE Transactions on Signal Processing*, 56(4):1466–1471, April 2008.
- [13] S. D. Babacan, L. Mancera, R. Molina, and A. K. Katsaggelos. Non convex priors in bayesian compressive sensing. In *Proceedings, European Signal Processing Conference*, 2009.
- [14] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Fast bayesian compressive sensing using laplace priors. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009.
- [15] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using laplace priors. *IEEE Transactions on Image Processing*, 19(1):53–63, January 2010.
- [16] Z. D. Bai and J. C. Fu. On the maximum–likelihood estimator for the location parameter of a cauchy distribution. *The Canadian Journal of Statistics*, 15(2):137–146, June 1987.
- [17] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak. Toeplitz-structured compressed sensing matrices. In *Proceedings, IEEE/SP 14th Workshop on Statistical Signal Processing*, August 2007.
- [18] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, July 2007.
- [19] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, December 2008.
- [20] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. To appear in *Foundations of Computational Mathematics*, 2008.
- [21] R.G. Baraniuk, E. Candès, M. Elad, and M. Yi. Applications of sparse representation and compressive sensing. *Proceedings of the IEEE*, 98(6):906–909, June 2010.
- [22] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, April 2010.

- [23] R.G. Baraniuk, V. Cevher, and M.B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, June 2010.
- [24] K. E. Barner and G. R. Arce. *Nonlinear Signal and Image Processing: Theory, Methods and Applications*. CRC Press, 2003.
- [25] K. E. Barner and T. C. Aysal. Polynomial weighted median filtering. *IEEE Transactions on Signal Processing*, 54(2):636–650, February 2006.
- [26] J. Bect, L. Blanc Feraud, G. Aubert, and A. Chambolle. *Lecture Notes in Computer Sciences 3024*, chapter A 11-unified variational framework for image restoration, pages 1–13. Springer Verlag, 2004.
- [27] R. Berinde, A. C. Gilbert, P. Indyk, and M. J. Strauss. Combining geometry and combinatorics: a unified approach to sparse signal recovery. Preprint, 2008.
- [28] D. P. Bertsekas. *Nonlinear Programming*. Athenea Scientific, Boston, 2nd ed. edition, 1999.
- [29] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B.*, 48(3):259–302, March 1986.
- [30] M. Blanco-Velasco, F. Cruz-Roldán, E. Moreno-Martínez, J. Godino-Llorente, and K. E. Barner. Embedded filter bank-based algorithm for ecg compression. *Signal Processing*, 88(6):1402 – 1412, 2008.
- [31] T. Blumensath. Compressed sensing with nonlinear observations. *Preprint*, 2011.
- [32] T. Blumensath and M. E. Davies. Iterative hard thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5):629 – 654, November 2008.
- [33] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265 – 274, November 2009.
- [34] T. Blumensath and M. E. Davies. Normalized iterative hard thresholding: guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):298–309, April 2010.

- [35] P. Boufounos, M. Duarte, and R. Baraniuk. Sparse signal reconstruction from noisy compressive measurements using cross validation. In *Proceedings, IEEE/SP 14th Workshop on Statistical Signal Processing*, Madison, WI, August 2007.
- [36] R.F. Brcich, D.R. Iskander, and A.M. Zoubir. The stability test for symmetric alpha-stable distributions. *Signal Processing, IEEE Transactions on*, 53(3):977–986, March 2005.
- [37] E. J. Candès. Compressive sampling. In *Proceedings, Int. Congress of Mathematics*, Madrid, Spain, August 2006.
- [38] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris, Series I*, pages 589–593, 2008.
- [39] E. J. Candès and P. A. Randall. Highly robust error correction by convex programming. 2006.
- [40] E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, April 2007.
- [41] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.
- [42] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, August 2006.
- [43] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.
- [44] E. J. Candès and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n . *Annal of statistics*, 2006.
- [45] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, December 2006.
- [46] E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, October 2009.

- [47] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, March 2008.
- [48] R. E. Carrillo, T. C. Aysal, and K. E. Barner. Generalized Cauchy distribution based robust estimation. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, April 2008.
- [49] R. E. Carrillo, T. C. Aysal, and K. E. Barner. A theoretical framework for problems requiring robust behavior. In *Proceedings, IEEE/EURASIP Workshop on Computational Advances in MultiSensor Adaptive Processing*, Aruba, Dutch Antilles, December 2009.
- [50] R. E. Carrillo, T. C. Aysal, and K. E. Barner. Bayesian compressed sensing using generalized Cauchy priors. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, March 2010.
- [51] R. E. Carrillo, T. C. Aysal, and K. E. Barner. A generalized Cauchy distribution framework for problems requiring robust behavior. *EURASIP Journal on Advances in Signal Processing*, 2010(Article ID 312989):19 pages, 2010.
- [52] R. E. Carrillo, T. C. Aysal, and K. E. Barner. Robust bayesian compressed sensing using generalized cauchy models. *IEEE Transactions on Image Processing*, July 2011. To be submitted.
- [53] R. E. Carrillo and K. E. Barner. Iteratively re-weighted least squares for sparse signal reconstruction from noisy measurements. In *Proceedings, Conference on Information Sciences and Systems*, Baltimore, MD, March 2009.
- [54] R. E. Carrillo and K. E. Barner. Lorentzian based iterative hard thresholding for compressed sensing. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.
- [55] R. E. Carrillo and K. E. Barner. Lorentzian iterative hard thresholding: Robust compressed sensing with prior information. *IEEE Transactions on Signal Processing*, July 2011. To be submitted.
- [56] R. E. Carrillo, K. E. Barner, and T. C. Aysal. Robust sampling and reconstruction methods for compressed sensing. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, April 2009.
- [57] R. E. Carrillo, K. E. Barner, and T. C. Aysal. Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):392–408, April 2010.

- [58] R. E. Carrillo, L. F. Polania, and K. E. Barner. Iterative algorithms for compressed sensing with partially known support. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, March 2010.
- [59] R. E. Carrillo, L. F. Polania, and K. E. Barner. Iterative hard thresholding for compressed sensing with partially known support. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.
- [60] R. M. Castro, J. Haupt, R. Nowak, and G. M. Raz. Finding needles in noisy haystacks. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, April 2008.
- [61] Volkan Cevher, Marco F. Duarte, Chinmay Hegde, and Richard G. Baraniuk. Sparse signal recovery using markov random fields. In *Proceedings of the Workshop on Neural Information Processing Systems (NIPS)*, Vancouver/Canada, Dec. 2008.
- [62] R. Chartrand, R. G. Baraniuk, Y. C. Eldar, M. A. T. Figueiredo, and J. Tanner. Introduction to the issue on compressive sensing. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):241–243, April 2010.
- [63] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(035020):1–14, 2008.
- [64] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their applications to nonlinear system identification. *Intl. J. Contr.*, 50(5):1873–1896, 1989.
- [65] S. Chen, D. L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- [66] J. F. Claerbout and F. Muir. Robust modelling with erratic data. *Geophys. Mag.*, 38(5):826–844, October 1973.
- [67] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22:211–231, 2009. Available online since Jul. 31, 2008.
- [68] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Applied Computational Harmonic Analysis*, 10(1):27–44, 2001.
- [69] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Journal on Multiscale Modeling and Simulation*, 4:11681700, November 2005.

- [70] W. Dai and O. Milenkovic. Weighted superimposed codes and constrained integer compressed sensing. Preprint, 2008.
- [71] I. Daubechies. *Ten lectures on wavelets*. CBS-NSF Regional Conferences in Applied Mathematics, 61, SIAM, 1992.
- [72] I. Daubechies, M. Defries, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.
- [73] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Gunturk. Iteratively re-weighted least squares for sparse approximation. *Communications on Pure and Applied Mathematics*, 63(1):1–38, October 2009.
- [74] R.N. Dave. Characterization and detection of noise in clustering. *Pattern Recognition Lett.*, 12(11):657–664, 1991.
- [75] R.N. Dave and R. Krishnapuram. Robust clustering methods: A unified view. *IEEE Trans. Fuzzy Systems*, 5(2):270–293, May 1997.
- [76] G. Davis, S. Mallat, and Z. Zhang. Adaptive time–frequency decompositions. *Opt. Eng.*, 33(7):2183–2191, July 1994.
- [77] R. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4-6):918–925, 2007.
- [78] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, May 1995.
- [79] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, September 2006.
- [80] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 56(6):797–829, June 2006.
- [81] D. L. Donoho. High–dimensional centrally–symetric polytopes with neighborliness proportional to dimesion. *Disc. Compt. Geometry*, 35(4):617–652, 2006.
- [82] D. L. Donoho and M. Elad. Optimally sparse representation from overcomplete dictionaries via l_1 norm minimization. In *Proc. Natl. Acad. Sci.*, USA, March 2002.

- [83] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, January 2006.
- [84] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, November 2001.
- [85] D. L. Donoho and J. Tanner. Neighborliness of randomly-projected simplices in high dimensions. In *Proc. National Academy of Sciences*, 2005.
- [86] D. L. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. Submitted to Journal of the AMS, 2008.
- [87] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. C. Daubechies. Data compression and harmonic analysis. *IEEE Transactions on Information Theory*, 44(6):2435–2476, October 1998.
- [88] D.L. Donoho and J. Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, 98(6):913–924, June 2010.
- [89] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressed sensing. *IEEE, Signal Processing Magazine*, 25(2):83–91, March 2008.
- [90] M. Duarte, C. Hegde, V. Cevher, and R. Baraniuk. Recovery of compressible signals in unions of subspaces. In *Proceedings, CISS 2009*, March 2009.
- [91] M. Duarte, M. Wakin, and R.G. Baraniuk. Fast reconstruction of piecewise smooth signals from random projections. In *Online Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, Rennes, France, 2005.
- [92] M. Duarte, M. Wakin, and R.G. Baraniuk. Wavelet-domain compressive signal reconstruction using a hidden markov tree model. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 5137–5140, 31 2008-April 4 2008.
- [93] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, April 2004.
- [94] M. Elad. Why simple shrinkage is still relevant for redundant representation. *IEEE Transactions on Information Theory*, 52(12):55591769, 2006.

- [95] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky. A wide-angle view at iterated shrinkage algorithms. In *SPIE (Wavelet XII)*, San Diego, CA, August 2007.
- [96] Y. Eldar. Compressed sensing of analog signals. Preprint, 2008.
- [97] Y.C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *Information Theory, IEEE Transactions on*, 55(11):5302–5316, Nov. 2009.
- [98] I. Esnaola, R. E. Carrillo, J. Garcia-Frias, and K. E. Barner. Orthogonal matching pursuit based recovery for correlated sources with partially disjoint supports. In *Proceedings, Conference on Information Sciences and Systems*, Princeton, NJ, March 2010.
- [99] P. Feng and Y. Bresler. Spectrum-blind minimum-rate sampling and reconstruction of multiband signals. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996.
- [100] M. Figueiredo and R. Nowak. An em algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906176, 2003.
- [101] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projections for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, December 2007.
- [102] J. H. Friedman and W. Stuetzle. Projection pursuit regressions. *Journal of the American Statistics Society*, 76:817–823, 1981.
- [103] J. Gao, R. E. Carrillo, and K. E. Barner. l_p metric based robust clustering. In *Proceedings, Conference on Information Sciences and Systems*, Baltimore, MD, March 2009.
- [104] J. Garcia-Frias and I. Esnaola. Exploiting prior knowledge in the recovery of signals from noisy random projections. In *Proceedings, IEEE Data Compression Conference*, Los Alamitos, CA, 2007.
- [105] A. C. Gilbert, S. Muthukrisan, and M. Strauss. Improved time bounds for near-optimal sparse fourier representation. In *Proceedings Wavelets XI SPIE Optics Photonics*, San Diego, CA, 2005.
- [106] J. G. Gonzales. *Robust Techniques for Wireless Communications in Non-Gaussian Environments*. Ph.D. dissertation, ECE Department, University of Delaware, 1997.

- [107] J. G. Gonzales and G. R. Arce. Optimality of the myriad filter in practical impulsive–noise environments. *IEEE Transactions on Signal Processing*, 49(2):438–441, February 2001.
- [108] J. G. Gonzales and G. R. Arce. Statistically–efficient filtering in impulsive environments: weighted myriad filters. *EURASIP Journal on Applied Signal Processing*, 2002(1):4–20, 2002.
- [109] J. G. Gonzales, J.L. Paredes, and G. R. Arce. Zero order statistics: a mathematical framework for the processing and characterization of very impulsive signals. *IEEE Transactions on Signal Processing*, 54(10):3839–3851, October 2006.
- [110] I. F. Gorodnitsky and B.D Rao. Sparse signal reconstruction from limited data using focuss: a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, March 1997.
- [111] R. Gribonval, V. Cevher, and M. Davies. Compressible priors for high-dimensional statistics. *Annals of Statistics*, 2011.
- [112] H. M. Hall. A new model for impulsive phenomena: application to atmospheric-noise communication channels. Technical report 3412 and 7050-7, Stanford Electronics Laboratories, Stanford University, Stanford, CA, 1966.
- [113] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust statistics: the approach based on influence functions*. New York: Wiley, 1986.
- [114] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge Mathematical Library (Reprint of the 1952 ed.), Cambridge: Cambridge University Press, 1988.
- [115] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, September 2006.
- [116] E. Hernandez and G. Weiss. *A first course on wavelets*. CRC Press, Inc., 1996.
- [117] K. K. Herrity, A. C. Gilbert, and J. A. Tropp. Sparse approximation via iterative thresholding. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, March 2006.
- [118] Huber. *Robust Statistics*. John Wiley & Sons, Inc., 1981.
- [119] L. Jacques. A short note on compressed sensing with partially known signal support. Technical Report, Université Catholique de Louvain, August 2009.

- [120] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, June 2008.
- [121] S. A. Kassam and H. V. Poor. Robust techniques for signal processing. *Proceedings of IEEE*, 73, March 1985.
- [122] S. J. Kim, k. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior point method for large-scale l_1 -regularized least squares problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, December 2007.
- [123] E. E. Kuruoglu. Signal processing with heavy-tailed distributions. *Signal Processing*, 82(12):1805 – 1806, Dec. 2002.
- [124] C. La and M. Do. Signal reconstruction using sparse tree representation. In *in Proc. Wavelets XI at SPIE Optics and Photonics*, 2005.
- [125] J. Laska, M. Davenport, and R. G. Baraniuk. Exact signal recovery from sparsely corrupted measurements through the pursuit of justice. In *Proceedings, IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2009.
- [126] J. Laska, S. Kirolos, M. Duarte, T. Ragheb, R. Baraniuk, and Y. Massoud. Theory and implementation of an analog-to-information converter using random demodulation. In *Proceedings, IEEE, Int. Symp. on Circuits and Systems*, New Orleans, LA, 2007.
- [127] Y.H. Ma, P.L. So, and E. Gunawan. Performance analysis of OFDM systems for broadband power line communications under impulsive noise and multipath effects. *IEEE Transactions on Power Delivery*, 20(2):674–682, April 2005.
- [128] S. Mallat and Z. Zhang. Matching pursuits with time frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):407–499, April 1993.
- [129] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, Inc., 1997.
- [130] S. Mendelson, A. Pajor, and N. Toczak-Jaegermann. Uniform uncertainty principle for bernoulli and sub-gaussian ensembles. *Constructive Approximation*, 28(3):277–289, December 2008.
- [131] D. Middleton. Statistical-physical models of electromagnetic interference. *IEEE Transactions on Electromagnetic Compatibility*, EMC-19(8):106–127, August 1977.

- [132] J. Miller and J. Thomas. Detectors for discrete-time signals in non-gaussian noise. *IEEE Transactions on Information Theory*, 8(2):241–250, March 1972.
- [133] H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. *IEEE Transactions on Signal Processing*, 57(1):289–301, January 2009.
- [134] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied Computational Harmonic Analysis*, 26(3):301–321, April 2008.
- [135] D. Needell and R. Vershynin. Uniform uncertainty principle and signal reconstruction via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, June 2008. Online.
- [136] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, 2006.
- [137] J. P. Nolan. *Stable Distributions: Models for Heavy Tailed Data*. Boston, MA: Birkhuser, 2005.
- [138] R. Nowak and M. Figueiredo. Fast wavelet-based image deconvolution using the em algorithm. In *Proceedings, 35th Asilomar Conference on Signals, Systems and Computers*, November 2001.
- [139] D. Omidiran and M. Wainwright. High-dimensional subset recovery in noise: Sparse measurements and statistical efficiency. In *Proceedings, IEEE, Int. Symp. on Information Theory*, Toronto, Canada, July 2008.
- [140] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao. Variational em algorithms for non-gaussian latent variable models. 2005.
- [141] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. Mc-Graw Hill, 1984.
- [142] J. Paredes and G. R. Arce. Compressive sensing signal reconstruction by weighted median regression estimates. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, March 2010.
- [143] Y. C. Patti, R. Rezaifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings, 27th Annual Asilomar Conf. Signals, Systems and Computers*, November 1993.

- [144] L. F. Polania, R. E. Carrillo, M. Blanco-Velazco, and K. E. Barner. Compressed sensing based method for ecg compression. In *Proceedings, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.
- [145] B. Popilka, S. Setzer, and G. Steidl. Signal recovery from incomplete measurements in the presence of outliers. *Inverse Problems and Imaging*, 1(4):661–672, November 2007.
- [146] S. Quian and D. Chen. Signal representation using adaptive normalized gaussian functions. *Signal Processing*, 36:329–355, 1994.
- [147] T. Ragheb, S. Kirolos, J. Laska, A. C. Gilbert, M. Strauss, R. Baraniuk, and Y. Massoud. Implementation models for analog-to-information conversion via random sampling. In *Midwest Symposium on Circuit and Systems*, 2007.
- [148] G. Reeves and M. Gastpar. Differences between observation and sampling error in sparse signal reconstruction. In *Proceedings of the 2007 IEEE Workshop on Statistical Signal Processing (SSP 2007)*, Madison, WI, August 2007.
- [149] G. Reeves and M. Gastpar. Sampling bounds for sparse support recovery in the presence of noise. In *Proceedings, IEEE, Int. Symp. on Information Theory*, Toronto, Canada, July 2008.
- [150] P. R. Rider. Generalized Cauchy distributions. *Annals of the Institute of Statistical Mathematics*, 9:215–223, 1957.
- [151] J. Romberg. Compressive sensing by random convolution. Preprint, 2008.
- [152] R. Saab and O. Yilmaz. Sparse recovery by non-convex optimization-instance optimality. *Applied and Computational Harmonic Analysis*, In Press, Corrected Proof, 2009.
- [153] F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7(4):1307–1330, 1986.
- [154] M. Shao and C.L. Nikias. Signal processing with fractional lower order moments: stable processes and their applications. *Proceedings of the IEEE*, 81(7):986–1010, Jul 1993.
- [155] J. Starck, M. Nguyen, and F. Murtagh. Wavelet and curvelet for image deconvolution: a combined approach. *Journal of Signal Processing*, 83(10):22791783, 2003.

- [156] D. S. Taubman and M. W. Marcellin. *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Kluwer, 2001.
- [157] V. Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3(1):33–107, July 2003.
- [158] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistics Society*, 58(1):267–288, 1996.
- [159] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
- [160] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2243, October 2004.
- [161] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, December 2007.
- [162] J.A. Tropp and S.J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, June 2010.
- [163] Y. Tsaig and D. L. Donoho. Extensions of compressed sensing. *Signal Processing*, 86(3):549–571, March 2006.
- [164] N. Vaswani and W. Lu. Modified-cs: Modifying compressive sensing for problems with partially known support. In *Proceedings, IEEE Int. Symp. Info. Theory*, 2009.
- [165] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing*, 50(6):1417–1428, June 2002.
- [166] M. Wakin. Manifold-based signal recovery and parameter estimation from compressive measurements. Preprint, 2008.
- [167] D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, April 2010.
- [168] J. Wright and Y. Ma. Dense error correction via l_1 minimization. Preprint, 2008.
- [169] M. Yang and K. Wu. A similarity-based robust clustering method. *IEEE Trans. Pattern Anal. Machine Intell.*, 26:434–448, Apr. 2004.

- [170] L. Yin, R. Yang, M. Gabbouj, and Y. Neuvo. Weighted median filters: A tutorial. *IEEE Transactions on Circuits and Systems*, 41, May 1996.
- [171] M. Zimmerman and K. Dostert. Analysis and modeling of impulsive noise in broadband power line communications. *IEEE Transactions on Electromagnetic Compatibility*, 44(1):249–558, February 2002.
- [172] V. Zolotarev. *One-Dimensional Stable Distributions*. Providence R.I.: American Mathematical Society, 1986.

Appendices

Appendix A

PROOF OF LEMMA 1: STATISTICAL RELATION BETWEEN GGD AND GCD RANDOM VARIABLES

Let X be the RV formed as the ratio of two RVs, U and V : $X = U/V$. In the case where U and V are independent, the PDF of the RV X , $f_X(\cdot)$, is given by [141]

$$f_X(x) = \int_{-\infty}^{\infty} |v| f_U(xv) f_V(v) dv, \quad (\text{A.1})$$

where $f_U(\cdot)$ and $f_V(\cdot)$ denote the PDFs of U and V , respectively. Replacing the GGD in (A.1) and manipulating the obtained expression yields

$$f_X(x) = \mathcal{C}(\alpha_U, \beta) \mathcal{C}(\alpha_V, \beta) \int_{v=-\infty}^{\infty} |v| \exp \left\{ - \left(\frac{|xv|}{\alpha_U} \right)^\beta \right\} \exp \left\{ - \left(\frac{|v|}{\alpha_V} \right)^\beta \right\} dv \quad (\text{A.2})$$

where $\mathcal{C}(\alpha, \beta) \triangleq \beta / (2\alpha\Gamma(1/\beta))$. Noting that $|ab| = |a||b|$ and dividing the integral gives

$$f_X(x) = \mathcal{C}(\alpha_U, \beta) \mathcal{C}(\alpha_V, \beta) \left[\int_{v>0} v \exp \left\{ -v^\beta \mathcal{K}(\alpha_U, \alpha_V, \beta, x) \right\} dv \right. \quad (\text{A.3}) \\ \left. - \int_{v\leq 0} v \exp \left\{ -(-v)^\beta \mathcal{K}(\alpha_U, \alpha_V, \beta, x) \right\} dv \right]$$

where

$$\mathcal{K}(\alpha_U, \alpha_V, \beta, x) \triangleq \frac{|x|^\beta}{\alpha_U^\beta} + \frac{1}{\alpha_V^\beta}.$$

Consider first

$$\mathcal{I}_1(v) \triangleq \int_{v>0} v \exp \{-v^\beta \mathcal{K}(\alpha_U, \alpha_V, \beta, x)\} dv. \quad (\text{A.4})$$

Letting $z = v^\beta \mathcal{K}(\alpha_U, \alpha_V, \beta, x)$, after some manipulations, yields

$$\mathcal{I}_1(v) = \frac{1}{\beta \mathcal{K}^{\frac{2}{\beta}}(\alpha_U, \alpha_V, \beta, x)} \int_{z>0} z^{\frac{2}{\beta}-1} \exp(-z) dz. \quad (\text{A.5})$$

Noting that

$$\int_{z>0} z^{\frac{2}{\beta}-1} \exp(-z) dz = \Gamma\left(\frac{2}{\beta}\right)$$

gives

$$\mathcal{I}_1(v) = \frac{1}{\beta \mathcal{K}^{\frac{2}{\beta}}(\alpha_U, \alpha_V, \beta, x)} \Gamma\left(\frac{2}{\beta}\right). \quad (\text{A.6})$$

Consider next

$$\mathcal{I}_2(v) \triangleq \int_{v \leq 0} v \exp \{-(-v)^\beta \mathcal{K}(\alpha_U, \alpha_V, \beta, x)\} dv. \quad (\text{A.7})$$

Letting $w = -v$, it is easy to see that $\mathcal{I}_2(v) = -\mathcal{I}_1(v)$, thus $\mathcal{I}_1(v) - \mathcal{I}_2(v) = 2\mathcal{I}_1(v)$.

Thus,

$$f_X(x) = \mathcal{C}(\alpha_U, \beta) \mathcal{C}(\alpha_V, \beta) 2\mathcal{I}_1 \quad (\text{A.8})$$

gives the desired result after substituting the corresponding expressions and letting $\alpha_U/\alpha_V = \nu$ and $\beta = \lambda$.

Appendix B

PROOF OF PROPOSITION 1: PROPERTIES OF THE M-GC COST FUNCTION

1. Differentiating $Q(\theta)$ yields

$$Q'(\theta) = \sum_{i=1}^N \frac{-p|x(i) - \theta|^{p-1} \text{sgn}(x(i) - \theta)}{\sigma^p + |x(i) - \theta|^p}. \quad (\text{B.1})$$

For $\theta < x_{[1]}$, $\text{sgn}(x(i) - \theta) = 1 \forall i$. Then $Q'(\theta) < 0$, which implies that $Q(\theta)$ is strictly decreasing in that interval. Similarly for $\theta > x_{[N]}$, $\text{sgn}(x(i) - \theta) = -1 \forall i$ and $Q'(\theta) > 0$, showing that the function is strictly increasing in that interval.

2. From 1) we see that $Q'(\theta) \neq 0$ if $\theta \notin [x_{[1]}, x_{[N]}]$ then all local extrema of $Q(\theta)$ lie in the interval $[x_{[1]}, x_{[N]}]$.
3. Let $x_{[k]} < \theta < x_{[k+1]}$ for any $k \in 1, 2, \dots, N - 1$. Then the objective function $Q(\theta)$ becomes

$$Q(\theta) = \sum_{i=1}^k \log\{\sigma^p + (\theta - x(i))^p\} + \sum_{i=k+1}^N \log\{\sigma^p + (x(i) - \theta)^p\}. \quad (\text{B.2})$$

The second derivative with respect to θ is

$$\begin{aligned}
Q''(\theta) &= \sum_{i=1}^k \frac{p(p-1)(\theta - x(i))^{p-2}\sigma^p - p(\theta - x(i))^{2p-2}}{(\sigma^p + (\theta - x(i))^p)^2} \\
&+ \sum_{i=k+1}^N \frac{p(p-1)(x(i) - \theta)^{p-2}\sigma^p - p(x(i) - \theta)^{2p-2}}{(\sigma^p + (x(i) - \theta)^p)^2}.
\end{aligned} \tag{B.3}$$

From (B.3) it can be seen that if $0 < p \leq 1$ then $Q''(\theta) < 0$ for $x_{[k]} < \theta < x_{[k+1]}$, therefore $Q(\theta)$ is concave in the intervals $I_k = (x_{[k]}, x_{[k+1]})$, $k \in 1, 2, \dots, N-1$. If all the extrema points lie in $[x_{[1]}, x_{[N]}]$, the function is concave in I_k and since the function is not differentiable in the input samples $\{x(i)\}_{i=1}^N$ (critical points) then the only possible local minimums of the objective function are the input samples.

4. Consider the i -th term in $Q(\theta)$ and define

$$q_i(\theta) = \log\{\sigma^p + |x(i) - \theta|^p\}. \tag{B.4}$$

Clearly for each $q_i(\theta)$ there exists a unique minima in $\theta = x(i)$. Also, it can be easily shown that $q_i(\theta)$ is convex in the interval $[x(i) - a, x(i) + a]$, where $a = \sigma(p-1)^{\frac{1}{p}}$, and concave outside this interval (for $1 < p \leq 2$). The proof of this statement is divided in two parts. First we consider the case when $N = 2$ and show that there exist at most $2N - 1 (= 3)$ local extrema for this case. Then by induction we generalize this result for any N .

Let $N = 2$. If $|x_{[2]} - x_{[1]}| < a$ the cost function is convex in the interval $[x_{[1]}, x_{[2]}]$ since is the sum of two convex functions (in that interval). Thus, $Q(\theta)$ has a unique minimizer. Now if $|x_{[2]} - x_{[1]}| \geq a$ the cost function has at most one inflexion point (local maxima) between $(x_{[1]}, x_{[2]})$ and at most two local minimas in the neighborhood of $x_{[1]}$ and $x_{[2]}$ since $q_i(\theta)$, $i = 1, 2$, are

concave outside the interval $[x_{[i]} - a, x_{[i]} + a]$. Then, for $N = 2$ we have at most $2N - 1 = 3$ local extrema points.

Suppose we have $N = K$ samples. If $|x_{[K]} - x_{[1]}| < a$ the cost function is convex in the interval $[x_{[1]}, x_{[K]}]$ since is the sum of convex functions (in that interval), and it has only one global minima. Now suppose that $|x_{[K]} - x_{[1]}| \geq a$ and also suppose that there are at most $2K - 1$ local extrema points. Let $x(K + 1)$ be a new sample in the data set and without loss of generality assume that $x(K + 1) > x_{[K]}$.

If $|x(K - 1) - x_{[K]}| < a$ the new sample will not add a new extrema point to the cost function, due to convexity of $q_{K+1}(\theta)$ for the interval $[x(K + 1) - a, x(K + 1) + a]$ and the fact that $Q(\theta)$ is strictly increasing for $\theta > x_{[K]}$. If $|x(K - 1) - x_{[K]}| \geq a$ the new sample will add at most two local extrema points (one local maxima and one local minima) in the interval $(x_{[K]}, x(K + 1))$. The local maxima is an inflexion point between $(x_{[K]}, x(K + 1))$ and the local minima is in the neighborhood of $x(K + 1)$. Therefore, the total number of extrema points for $N = K + 1$ is at most $2K - 1 + 2 = 2(K + 1) - 1$, which is the claim of the statement. This concludes the proof.

Appendix C

ITERATIVE ALGORITHMS FOR CS WITH PARTIALLY KNOWN SUPPORT

In this appendix we describe extensions of three iterative algorithms to incorporate the partially known support in to the iterative process. The iterative algorithms are: OMP, CoSaMP and RWLS- SL_0 . Results of these approaches are presented in [58].

C.1 OMP

OMP is an iterative greedy algorithm for sparse signal recovery [161]. At each iteration, we choose the column of Φ that is most strongly correlated with the remaining part of the signal y . Then we subtract off its contribution to the measurement vector and iterate on the residual.

Since the algorithm needs to determine which columns of Φ participate in the measurement vector, it is natural to think of the introduction of partially known support ideas to enhance its recovery performance. Thus, the partially known support gives *a priori* information about some of the columns that should be selected. This piece of information modifies the initialization of the algorithm because we need to subtract off the contribution of these columns to the measurement vector before starting the iteration. Therefore, the residual needs to be initialized as

$$r = y - \Phi_{T_0}(\Phi_{T_0}^\dagger y), \tag{C.1}$$

where T_0 is the partially known support and the initial support of the signal at $t = 0$.

The algorithm terminates when the L_2 norm of the residual falls below a selected approximation error bound. All the steps in the iteration remain the same as in OMP. To summarize, the final algorithm is depicted in Algorithm 7.

Algorithm 7 OMP Algorithm with partially known support

Require: CS matrix Φ , measurements y and partial known support T_0 .

- 1: Initialize $i = 0$, $\hat{x}_0 = \Phi_{T_0}^\dagger y$ and $r = y - \Phi_{T_0} \hat{x}_0$.
 - 2: **while** halting criterion **do**
 - 3: $i \leftarrow i + 1$.
 - 4: $e \leftarrow \Phi^T r$.
 - 5: $\Omega \leftarrow \arg \max_j |e(j)|$
 - 6: $T \leftarrow \Omega \cup \text{supp}(\hat{x}_{i-1})$
 - 7: $\hat{x}_i \leftarrow \Phi_T^\dagger y$
 - 8: $r \leftarrow y - \Phi_T \hat{x}_i$
 - 9: **end while**
 - 10: **return** $x \leftarrow \hat{x}_i$
-

C.2 CoSaMP

Compressive Sampling Matching Pursuit (CoSaMP) is also a greedy algorithm [134]. Then, as in the orthogonal matching pursuit case, the ideas of partially known support can be incorporated and the initialization needs to be modified in a similar way as to that for OMP. Thus the residual is calculated by subtracting the contribution of the first estimate. Additionally, we calculate the first estimate of the signal by solving a least squares problem using Φ_{T_0} .

In one step of the iteration process, CoSaMP identifies the $2s$ largest components of the signal proxy. Since we already know a subset of the support, we just need to identify the $2(s - |T_0|)$ largest components instead.

CoSaMP prunes the signal to be s -sparse. In order to do that and include the a priori known information, an approximation to the signal is formed at each

iteration by selecting the largest coordinates and the ones that correspond to the partially known support. The rest of the algorithm remains the same as CoSaMP. The entire algorithm is specified in Algorithm 8.

Algorithm 8 CoSaMP Algorithm with partially known support

Require: CS matrix Φ , measurements y , sparsity level s and partial known support T_0 .

- 1: Initialize $\hat{x}_0|_{T_0} = \Phi_{T_0}^\dagger y$, $\hat{x}_0|_{T_0^c} = 0$, $r = y - \Phi_{T_0} \hat{x}_0|_{T_0}$, $K = s - |T_0|$ and $i = 0$.
 - 2: **while** halting criterion false **do**
 - 3: $i \leftarrow i + 1$.
 - 4: $e \leftarrow \Phi^T r$.
 - 5: $\Omega \leftarrow \text{supp}(e_{2K})$
 - 6: $T \leftarrow \Omega \cup \text{supp}(\hat{x}_{i-1})$
 - 7: $b|_T \leftarrow \Phi_T^\dagger y$, $b|_{T^c} \leftarrow 0$
 - 8: $A|_{T_0^c} \leftarrow b|_{T_0^c}$, $A|_{T_0} \leftarrow 0$
 - 9: $\hat{x}_i \leftarrow A|_{(T_0 \cup \text{supp}(A_K))}$
 - 10: $r \leftarrow y - \Phi \hat{x}_i$
 - 11: **end while**
 - 12: **return** $x \leftarrow \hat{x}_i$
-

C.3 RWLS- S_{L_0}

As described in [53], the iterative reweighted least squares approach based on smooth approximation of the L_0 norm is an efficient method to reconstruct sparse signals. The following function, which converges pointwise to the L_0 norm as $\sigma \rightarrow 0$, was proposed in [53]:

$$F_\sigma(x) = \sum_{i=1}^n f_\sigma(x_i) = \sum_{i=1}^n \frac{|x_i|}{\sigma + |x_i|}. \quad (\text{C.2})$$

In order to find the sparsest possible signal estimate whose support contains T_0 , we propose to solve the following problem

$$\min_{x \in \mathbb{R}^n} \sum_{i \notin T_0} \frac{|x_i|}{\sigma + |x_i|} \text{ s.t. } \|y - \Phi x\|_2 \leq \epsilon. \quad (\text{C.3})$$

To solve the nonconvex optimization problem derived, an iterative re-weighted least squares approach, whose purpose is to encourage sparse solutions by giving a large weight to small components, was proposed in the paper. Since the objective is not convex and can have several local minima on the feasible set, a convex problem was introduced to be solved iteratively.

We propose to rewrite the solution of the problem at iteration t as

$$\hat{x}^{t+1} = W^t \Phi^T (\Phi W^t \Phi^T + \lambda I)^{-1} y,$$

where λ is a small regularization parameter set as some predefined $\lambda_{min} > 0$. We also need to rewrite the diagonal weighting matrix W_t such that its diagonal elements become

$$W_{ii}^t = (\sigma^t + |\hat{x}_i^t|)^2,$$

It is natural to think that the elements of the diagonal whose positions correspond to the partially known support should have a much greater value than the others. We set this value as one hundred times the largest element of the diagonal.

Appendix D

PROOF OF THEOREM 8: STABILITY OF THE LIHT-PKS ALGORITHM

Suppose $x \in \mathbb{R}^n$ and $T = \text{supp}(x)$, $|T| = s$ (s -sparse signal). If $T = T_0 \cup \Delta$, then $|\Delta| = s - k$ where $|T_0| = k$. Define

$$a^{(t)} = x^{(t)} + \Phi^T W_t (y - \Phi x^{(t)}). \quad (\text{D.1})$$

The update at each iteration $t + 1$ can be expressed as:

$$x^{(t+1)} = a_{T_0}^{(t)} + H_{s-k}(a_{T_0^c}^{(t)}) \quad (\text{D.2})$$

and the residual (reconstruction error) at iteration t is defined as $r^{(t)} = x - x^{(t)}$.

Define $T^{(t)} = \text{supp}(x^{(t)})$ and $U^{(t)} = \text{supp}\left(H_{s-k}(a_{T_0^c}^{(t)})\right)$. It can be easily checked for all t that $|\text{supp}(a_{T_0}^{(t)})| = k$, $|U^{(t)}| = s - k$ and $|T^{(t)}| = s$. Also define

$$B^{(t+1)} = T \cup T^{(t+1)} = T_0 \cup \Delta \cup U^{(t+1)}.$$

Then, the cardinality of the set $B^{(t+1)}$ is upper bounded by

$$|B^{(t+1)}| \leq |T_0| + |\Delta| + |U^{(t+1)}| = 2s - k.$$

The error $r^{(t)}$ is supported on $B^{(t+1)}$. Using the triangle inequality we have

$$\|x_{B^{(t+1)}} - x_{B^{(t+1)}}^{(t+1)}\|_2 \leq \|x_{B^{(t+1)}} - a_{B^{(t+1)}}^{(t+1)}\|_2 + \|x_{B^{(t+1)}}^{(t+1)} - a_{B^{(t+1)}}^{(t+1)}\|_2.$$

We start by bounding $\|x_{B^{(t+1)}}^{(t)} - a_{B^{(t+1)}}^{(t)}\|_2$. Remember that

$$x^{(t+1)} = x_{T_0}^{(t+1)} + x_{T_0^c}^{(t+1)}, \quad a^{(t+1)} = a_{T_0}^{(t+1)} + a_{T_0^c}^{(t+1)}.$$

By definition $x_{T_0}^{(t+1)} = a_{T_0}^{(t+1)}$. By the thresholding operator, $x_{T_0}^{(t+1)}$ is the best $(s-k)$ -term approximation of $a_{T_0}^{(t+1)}$. Then, $x^{(t+1)}$ is a better approximation to $a^{(t+1)}$ than x and we have

$$\|x_{B^{(t+1)}}^{(t+1)} - a_{B^{(t+1)}}^{(t+1)}\|_2 \leq \|x_{B^{(t+1)}} - a_{B^{(t+1)}}^{(t+1)}\|_2.$$

Therefore the error at iteration $t+1$ is bounded by

$$\|x_{B^{(t+1)}} - x_{B^{(t+1)}}^{(t+1)}\|_2 \leq 2\|x_{B^{(t+1)}} - a_{B^{(t+1)}}^{(t+1)}\|_2.$$

Rewrite (D.1) as

$$a^{(t+1)} = x^{(t)} + \Phi^T W_t \Phi x - \Phi^T W_t \Phi x^{(t)} + \Phi^T W_t z.$$

Denote Φ_Ω as the submatrix obtained by selecting the columns indicated by Ω . Then

$$a_{B^{(t+1)}}^{(t+1)} = x_{B^{(t+1)}}^{(t)} + \Phi_{B^{(t+1)}}^T W_t \Phi r^{(t)} + \Phi_{B^{(t+1)}}^T W_t z$$

and we can bound the estimation error as

$$\begin{aligned}
\|x_{B^{(t+1)}} - x_{B^{(t+1)}}^{(t+1)}\|_2 &\leq 2\|x_{B^{(t+1)}} - x_{B^{(t+1)}}^{(t)} - \Phi_{B^{(t+1)}}^T W_t \Phi r^{(t)} - \Phi_{B^{(t+1)}}^T W_t z\|_2 \\
&\leq 2\|r_{B^{(t+1)}}^{(t)} - \Phi_{B^{(t+1)}}^T W_t \Phi r^{(t)}\|_2 + 2\|\Phi_{B^{(t+1)}}^T W_t z\|_2 \\
&\leq 2\|(I - \Phi_{B^{(t+1)}}^T W_t \Phi_{B^{(t+1)}})r_{B^{(t+1)}}^{(t)} - \Phi_{B^{(t+1)}}^T W_t \Phi_{B^{(t)} \setminus B^{(t+1)}} r_{B^{(t)} \setminus B^{(t+1)}}^{(t)}\|_2 \\
&\quad + 2\|\Phi_{B^{(t+1)}}^T W_t z\|_2 \\
&\leq 2\|(I - \Phi_{B^{(t+1)}}^T W_t \Phi_{B^{(t+1)}})r_{B^{(t+1)}}^{(t)}\|_2 \\
&\quad + 2\|\Phi_{B^{(t+1)}}^T W_t \Phi_{B^{(t)} \setminus B^{(t+1)}} r_{B^{(t)} \setminus B^{(t+1)}}^{(t)}\|_2 + 2\|\Phi_{B^{(t+1)}}^T W_t z\|_2.
\end{aligned}$$

Now since $[W_t]_{i,i} \leq 1$, then the eigenvalues of $\Phi^T W_t \Phi$ are bounded above by the eigenvalues of $\Phi^T \Phi$, and therefore

$$\begin{aligned}
\|x_{B^{(t+1)}} - x_{B^{(t+1)}}^{(t+1)}\|_2 &\leq 2\|(I - \Phi_{B^{(t+1)}}^T \Phi_{B^{(t+1)}})r_{B^{(t+1)}}^{(t)}\|_2 \\
&\quad + 2\|\Phi_{B^{(t+1)}}^T \Phi_{B^{(t)} \setminus B^{(t+1)}} r_{B^{(t)} \setminus B^{(t+1)}}^{(t)}\|_2 + 2\|\Phi_{B^{(t+1)}}^T W_t z\|_2.
\end{aligned}$$

Notice that

$$\begin{aligned}
|B^{(t)} \cup B^{(t+1)}| &= |T_0 \cup \Delta \cup U^{(t+1)} \cup U^{(t)}| \\
&\leq |T_0| + |\Delta| + 2|U^{(t)}| = 3s - 2k.
\end{aligned}$$

Using basic properties of the restricted isometry constants (see Lemma 1 from [33])

and the fact that $\delta_{3s-2k} > \delta_{2s-k}$ we have the following. Define $\eta = 2\sqrt{1 + \delta_{2s-k}}$.

$$\begin{aligned}
\|x_{B^{(t+1)}} - x_{B^{(t+1)}}^{(t+1)}\|_2 &\leq 2\delta_{2s-k}\|r_{B^{(t+1)}}^{(t)}\|_2 + 2\delta_{3s-2k}\|r_{B^{(t)} \setminus B^{(t+1)}}^{(t)}\|_2 + \eta\|W_t z\|_2 \\
&\leq 2\delta_{3s-2k}(\|r_{B^{(t+1)}}^{(t)}\|_2 + \|r_{B^{(t)} \setminus B^{(t+1)}}^{(t)}\|_2) + \eta\|W_t z\|_2.
\end{aligned}$$

Since $B^{(t)} \setminus B^{(t+1)}$ and $B^{(t+1)}$ are disjoint sets we have $\|r_{B^{(t+1)}}^{(t)}\|_2 + \|r_{B^{(t)} \setminus B^{(t+1)}}^{(t)}\|_2 \leq$

$\sqrt{2}\|r_{B^{(t)} \cup B^{(t+1)}}^{(t)}\|_2$. Then, the estimation error at iteration $t + 1$ is bounden by

$$\|r^{(t+1)}\|_2 \leq \sqrt{8}\delta_{3s-2k}\|r^{(t)}\|_2 + \eta\|W_t z\|_2.$$

This is a recursive error bound. Define $\alpha = \sqrt{8}\delta_{3s-2k}$ and assume $x^{(0)} = 0$. Then

$$\|r^{(t)}\|_2 \leq \alpha^t \|x\|_2 + \eta \|W_t z\|_2 \sum_{j=0}^{t-1} \alpha^j. \quad (\text{D.3})$$

We need $\alpha = \sqrt{8}\delta_{3s-2k} < 1$ for the series in (D.3) to converge. For faster convergence and better stability we restrict $\sqrt{8}\delta_{3s-2k} < 1/2$, which yields the sufficient condition in Theorem 8. Now we just need to bound $\|z\|_2$. Note that $W_t(i, i) \leq 1$, which implies that

$$\|W_t z\|_2 \leq \|z\|_2 \leq \gamma \sqrt{m(e^\epsilon - 1)},$$

where the second inequality follows from Lemma 1 in [57].