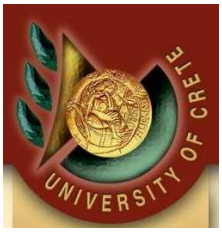


University of Crete
Department of Computer Science

FO.R.T.H.
Institute of Computer Science

Bayesian Compressed Sensing using Alpha-Stable Distributions

Georgios Tzagkarakis

A thesis submitted for the degree of
Doctor of Philosophy

Heraklion, November 2009

© Copyright by Georgios Tzagkarakis, 2009

All Rights Reserved

University of Crete
Department of Computer Science

Bayesian Compressed Sensing using Alpha-Stable Distributions

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

November 6, 2009

Graduate Committee Approval

Author:

Georgios Tzagkarakis, Dept. of Computer Science

Date	Panagiotis Tsakalides (Associate Professor–Advisor) Dept. of Computer Sc., Univ. of Crete
Date	Apostolos Traganitis (Professor), Dept. of Computer Sc., Univ. of Crete
Date	Yannis Stylianou (Associate Professor), Dept. of Computer Sc., Univ. of Crete
Date	Maria Papadopouli (Assistant Professor), Dept. of Computer Sc., Univ. of Crete
Date	Michael Taroudakis (Professor), Dept. of Mathematics, Univ. of Crete
Date	Georgios Tsihrintzis (Professor), Dept. of Informatics, Univ. of Piraeus
Date	Alin Achim (Lecturer), Dept. of Electrical & Electronic Eng., Univ. of Bristol

Approved by:

Dimitris Plexousakis
Chairman of the Department

To all those who guided and taught me all these years!

Acknowledgements

This dissertation is the apex of a long-lasting enterprizing journey and not simply the result of a working experience as a graduate student. I owe my gratitude to all the people who inspired me and made this thesis possible and because of whom my academic experience has been one that I will cherish forever. Starting from my school teachers, to my professors at the departments of Mathematics and Computer Science of the University of Crete (UOC).

Special thanks go to Prof. Panagiotis Tsakalides, my advisor, not only for his guidance and encouragement during this work, but also for giving me the freedom to choose and explore several research directions and for helping me to unfold my research style throughout the last 7 years of collaboration.

Thanks are due to Profs.: Apostolos Traganitis, Yannis Stylianou and Maria Papadopouli (Dept. of Computer Science–UOC), Prof. Michael Taroudakis (Dept. of Mathematics–UOC), Prof. George Tsihrintzis (Dept. of Informatics–Univ. of Piraeus) and last but not least, Prof. Alin Achim (Dept. of Electrical & Electronic Eng.–Univ. of Bristol) for agreeing to serve on my thesis committee and for sparing their valuable time reviewing the manuscript. My colleagues at the Telecommunications & Networks laboratory have enriched my graduate life in many ways and deserve a special mention.

This work was supported by a graduate fellowship from the Institute of Computer Science (ICS) of the Foundation for Research and Technology Hellas (FORTH), as well as by the Greek General Secretariat for Research and Technology under Program ΠΕΝΕΔ-Code 03ΕΔ69 and by the Marie Curie TOK-DEV "ASPIRE" grant (MTKD-CT-2005-029791) within the 6^{th} European Community Framework Program, which I also truly acknowledge for providing the necessary technical equipment. I would also like to acknowledge help and support from the secretariat and the administrators of the department of Computer Science and ICS-FORTH.

Finally, I would like to thank my family, for providing me with the ideal environment to grow up in, and exhorting me in changing for the better. Their constant support and encouragement have really brought me here. It is impossible to remember all and I apologize to those I've unintentionally left out.

This dissertation is part of the 03EΔ69 *research project, implemented within the framework of the "Reinforcement Program of Human Research Manpower"* (ΠΕΝΕΔ) *and co-financed by National and Community Funds (*25% *from the Greek Ministry of Development - General Secretariat of Research and Technology and* 75% *from E.U. - European Social Fund)*

Abstract

Bayesian Compressed Sensing using Alpha-Stable Distributions

Georgios Tzagkarakis
University of Crete
Department of Computer Science
Doctor of Philosophy, 2009

During the last decades, information is being gathered and processed at an explosive rate. This fact gives rise to a very important issue, that is, how to effectively and precisely describe the information content of a given source signal or an ensemble of source signals, such that it can be stored, processed or transmitted by taking into consideration the limitations and capabilities of the several digital devices.

One of the fundamental principles of signal processing for decades is the *Nyquist-Shannon sampling theorem*, which states that the minimum number of samples needed to reconstruct a signal without error is dictated by its bandwidth. However, there are many cases in our everyday life in which sampling at the Nyquist rate results in too many data and thus, demanding an increased processing power, as well as storage requirements. A mathematical theory that emerged recently presents the background for developing a novel sensing/sampling paradigm that goes against the common tenet in data acquisition. *Compressed sensing* (CS), also known as compressive sensing, compressive sampling and sparse sampling, is a technique for acquiring and reconstructing a signal utilizing the prior knowledge that it is *sparse* or *compressible*, which provides a stricter sampling condition yielding a sub-Nyquist sampling criterion. Sparsity expresses the fact that the "information rate" of a continuous-time signal may be much smaller than suggested by its bandwidth via the Nyquist's theorem, or that a discrete-time signal depends on a number of degrees of freedom which is comparably much smaller than its (finite) length.

Several deterministic and probabilistic approaches have been proposed over the last years confronting the problem of sparse signal reconstruction from distinct viewpoints. The majority of these methods relies on solving constrained-based optimization problems by employing several vector norms in the design of appropriate objective functions. Only recently the problem of CS reconstruction has been studied in a probabilistic (Bayesian) framework, resulting in several advantages when compared with the norm-based techniques. However, both of these classes of algorithms are based on a Gaussian assumption for the characterization of the statistics of the sparse signal.

This thesis introduces the class of heavy-tailed distributions and particularly the family of *alpha-Stable* distributions as a suitable modeling tool for designing efficient CS reconstruction algorithms exploiting the sparsity of the received signal in an appropriate transform domain. More specifically, the first of the proposed methods exploits the prior knowledge for a sparse coefficient vector by modeling the statistics of its components using a Gaussian Scale Mixture

(GSM). Then, the reconstruction of the sparse signal is reduced to the problem of estimating the parameters of the GSM model, which is in turn carried out by developing a Bayesian technique. Furthermore, there are applications, for instance in the case of a sensor network, where the acquisition process results in a set of multiple observations of the unknown sparse signal. For this purpose, we extend the previous method in order to take into account the fact that the set of multiple observations is characterized by a common sparsity structure with high probability, yielding an efficient CS method amenable to a distributed implementation.

There are also real-world environments, such as in underwater acoustics or in telecommunications, where the signal and/or the noise present a highly impulsive behavior and thus, resulting in an even increased sparsity. The second method proposed in this thesis, which is also developed in a Bayesian framework, reconstructs signals corrupted by a highly impulsive noise component. This is done by modeling the statistical behavior of the sparse vector with a Cauchy distribution, which is also a member of the family of alpha-Stable distributions. The estimation of model parameters is performed by employing a well-known tree structure, which is a common approach in several Bayesian Learning tasks, such as the *model selection* as is the case here.

Finally, a third method is proposed that generalizes the other two in the sense that it is developed for an arbitrary alpha-Stable distribution. The efficiency of a CS reconstruction method does not only depend on the inversion (decoding) method by itself, but also on a suitable encoding part (measurement matrix) that embeds all the significant information of the few large-amplitude transform coefficients into a measurement vector. So, it is shown first that such a measurement matrix is constructed exploiting the accuracy of an alpha-Stable distribution in modeling the highly impulsive (sparse) behavior of the sparse (transform) coefficients vector. The proposed CS algorithm for estimating a sparse vector proceeds by solving iteratively a constrained optimization problem using the duality theory and the method of subgradients. However, since the family of alpha-Stable distributions lacks finite variance, we introduce a modified Lagrangian function which takes into account the true non-Gaussian behavior as expressed by the so called *Fractional Lower-Order Moments*. In addition, it is shown that the objective function and the constraints are separable and thus, this algorithm is amenable to a distributed implementation from the nodes of a sensor network.

We also illustrate the increased performance of the proposed CS algorithms based on heavy-tailed models when compared with the performance of recently introduced state-of-the-art CS reconstruction techniques, by applying them in a series of experiments for reconstructing simulated signals or for solving the problem of Direction-of-Arrival (DOA) estimation, as well as for recovering real image data from their corresponding transform coefficients.

Περίληψη

Μπεϋζιανή Συμπιεστική Δειγματοληψία με χρήση Άλφα-Ευσταθών Κατανομών

Γεώργιος Τζαγκαράκης
Πανεπιστήμιο Κρήτης
Τμήμα Επιστήμης Υπολογιστών
Διδακτορική Διατριβή, 2009

Κατά τη διάρκεια των τελευταίων ετών, η συλλογή και επεξεργασία της πληροφορίας πραγματοποιείται με πολύ ταχείς ρυθμούς. Αυτό το γεγονός θέτει ένα κρίσιμο ζήτημα αναφορικά με το πώς θα μπορέσει να περιγραφεί η πληροφορία που ενυπάρχει σε ένα σήμα ή σε ένα σύνολο σημάτων με αποδοτικό και ακριβή τρόπο, έτσι ώστε στη συνέχεια να αποθηκευτεί, επεξεργαστεί ή μεταδοθεί λαμβάνοντας υπόψη τις περιορισμένες δυνατότητες που έχουν οι διάφορες ψηφιακές συσκευές, όπως για παράδειγμα η περιορισμένη ενέργεια.

Μία από τις θεμελιώδεις αρχές στην επεξεργασία σήματος εδώ και δεκαετίες είναι το *θεώρημα δειγματοληψίας των Nyquist-Shannon*, σύμφωνα με το οποίο το ελάχιστο πλήθος δειγμάτων που απαιτείται για την ανακατασκευή ενός σήματος χωρίς σφάλμα καθορίζεται από το εύρος του φάσματός του. Εντούτοις, υπάρχουν πολλές περιπτώσεις στην καθημερινότητα στις οποίες όταν η δειγματοληψία πραγματοποιείται σύμφωνα με το ρυθμό Nyquist καταλήγει σε ένα υπερβολικά μεγάλο πλήθος δεδομένων και επομένως απαιτεί αυξημένη επεξεργαστική ισχύ, όπως επίσης και αποθηκευτικό χώρο. Μία μαθηματική θεωρία που ήρθε στο προσκήνιο πρόσφατα παρέχει το πλαίσιο για την ανάπτυξη ενός καινοτόμου προτύπου δειγματοληψίας, το οποίο αντικρούει τη συνήθη τάση στη λήψη δεδομένων (ρυθμός Nyquist). Η *συμπιεστική δειγματοληψία* (*compressed sensing*), γνωστή επίσης και ως *αραιή δειγματοληψία*, είναι μία τεχνική για τη λήψη και ανακατασκευή ενός σήματος που εκμεταλλεύεται την εκ των προτέρων γνώση ότι το σήμα είναι *αραιό ή συμπιεστό*, παρέχοντας μία πιο αυστηρή συνθήκη δειγματοληψίας που οδηγεί σε ρυθμούς μικρότερους από αυτόν του Nyquist. Η αραιότητα εκφράζει το γεγονός ότι ο *ρυθμός πληροφορίας* ενός σήματος συνεχούς χρόνου είναι πιθανόν πολύ μικρότερος από αυτόν που υποδεικνύει το εύρος του φάσματός του μέσω του θεωρήματος των Nyquist-Shannon, ή το ότι ένα σήμα διακριτού χρόνου εξαρτάται από ένα πλήθος βαθμών ελευθερίας το οποίο είναι πολύ μικρότερο σε σύγκριση με το (πεπερασμένο) μήκος του.

Διάφορες ντετερμινιστικές και πιθανοκρατικές προσεγγίσεις έχουν προταθεί τα τελευταία χρόνια, οι οποίες αντιμετωπίζουν το πρόβλημα της ανακατασκευής ενός αραιού σήματος από διαφορετικές οπτικές. Η πλειοψηφία των μεθόδων αυτών βασίζεται στην επίλυση ενός προβλήματος βελτιστοποίησης με περιορισμούς χρησιμοποιώντας διάφορες διανυσματικές νόρμες για τη σχεδίαση κατάλληλων αντικειμενικών συναρτήσεων. Μόλις πρόσφατα το πρόβλημα της ανακατασκευής ενός αραιού σήματος με χρήση συμπιεστικής δειγματοληψίας άρχισε να μελετάται σε ένα Μπεϋζιανό πλαίσιο, προσφέροντας διάφορα πλεονεκτήματα σε σχέση με τις προηγούμενες τεχνικές. Παρόλα αυτά και οι δύο αυτές κατηγορίες μεθόδων βασίζονται ως επί το πλείστον στην Γκαουσιανή υπόθεση για το στατιστικό χαρακτηρισμό του αραιού σήματος.

Η παρούσα εργασία εισάγει μια κλάση κατανομών με *βαριές ουρές*, πιο συγκεκριμένα την οικογένεια των *Άλφα-Ευσταθών* κατανομών (*alpha-Stable* distributions) ως ένα κατάλληλο μοντέλο για τη σχεδίαση και υλοποίηση αποδοτικών αλγορίθμων ανακατασκευής αραιού σήματος με χρήση συμπιεστικής δειγματοληψίας, οι οποίοι εκμεταλλεύονται την αραιότητα του λαμβανόμενου σήματος σε κάποιο πεδίο μετασχηματισμού. Ειδικότερα, η πρώτη από τις προτεινόμενες μεθόδους εκμεταλλεύεται την εκ των προτέρων γνώση ότι το διάνυσμα των συντελεστών του μετασχηματισμένου σήματος είναι αραιό μοντελοποιώντας τη στατιστική συμπεριφορά των συνιστωσών του με ένα μείγμα Γκαουσιανών (*Gaussian Scale Mixture (GSM)*). Κατόπιν, η ανακατασκευή του αρχικού σήματος ανάγεται σε πρόβλημα εκτίμησης των παραμέτρων του GSM μοντέλου, το οποίο επιλύεται με την προτεινόμενη Μπεϋζιανή μέθοδο. Επιπλέον, υπάρχουν εφαρμογές, για παράδειγμα στην περίπτωση ενός δικτύου αισθητήρων, όπου η διαδικασία λήψης καταλήγει σε ένα σύνολο πολλαπλών παρατηρήσεων του αρχικού σήματος. Για το σκοπό αυτό επεκτείνουμε την πρώτη μέθοδο ώστε να λαμβάνει υπόψη το γεγονός ότι το σύνολο των πολλαπλών παρατηρήσεων χαρακτηρίζεται, με μεγάλη πιθανότητα, από μία *κοινή αραιή δομή* οδηγώντας σε μια αποδοτική μέθοδο η οποία επιδέχεται *κατανεμημένη υλοποίηση*, κάτι πολύ σημαντικό σε ένα δίκτυο αισθητήρων.

Υπάρχουν επίσης πραγματικά περιβάλλοντα, όπως για παράδειγμα στην υποβρύχια ακουστική και τις τηλεπικοινωνίες, όπου το σήμα και/ή ο θόρυβος εμφανίζουν μια υψηλή κρουστικότητα και επομένως καταλήγουν σε ακόμα πιο αυξημένη αραιότητα. Η δεύτερη μέθοδος που προτείνεται σε αυτή την εργασία, η οποία επίσης αναπτύσσεται σε ένα Μπεϋζιανό πλαίσιο, ανακατασκευάζει κρουστικά σήματα στα οποία έχει προστεθεί θόρυβος με επίσης υψηλά κρουστική συμπεριφορά. Αυτό αντιμετωπίζεται μοντελοποιώντας το αραιό, κρουστικό σήμα με μία κατανομή Cauchy, η οποία ανήκει στην οικογένεια των *Άλφα-Ευσταθών* κατανομών. Η εκτίμηση των παραμέτρων του μοντέλου πραγματοποιείται χρησιμοποιώντας μία δενδρική δομή, η οποία συναντάται συχνά σε προβλήματα Μπεϋζιανής Μάθησης (Bayesian Learning), όπως για παράδειγμα στο πρόβλημα της *επιλογής μοντέλου* που είναι ουσιαστικά και η περίπτωσή μας εδώ.

Τέλος, μία τρίτη μέθοδος προτείνεται η οποία αποτελεί γενίκευση των προηγούμενων, με την έννοια ότι μπορεί να χρησιμοποιηθεί στην περίπτωση μοντελοποίησης χρησιμοποιώντας οποιαδήποτε *Άλφα-Ευσταθή* κατανομή. Η αποδοτικότητα ενός αλγορίθμου ανακατασκευής αραιού σήματος με χρήση συμπιεστικής δειγματοληψίας δεν εξαρτάται μόνο από τη μέθοδο (αποκωδικοποίηση) αυτή καθεαυτή, αλλά επίσης και από τον τρόπο λήψης των μετρήσεων (κωδικοποίηση). Για το σκοπό αυτό σημαντική είναι η κατασκευή ενός κατάλληλου πίνακα μετρήσεων (*measurement matrix*) ο οποίος θα εμφυτεύσει όλη τη σημαντική πληροφορία των λίγων, σημαντικών συντελεστών του (μετασχηματισμένου) αραιού σήματος σε ένα διάνυσμα μετρήσεων. Έτσι, αποδεικνύεται πρώτα ότι ένας τέτοιος πίνακας μετρήσεων κατασκευάζεται βασιζόμενος στην ακρίβεια με την οποία η *Άλφα-Ευσταθής* κατανομή μπορεί να μοντελοποιήσει την κρουστική συμπεριφορά του αραιού διανύσματος. Στη συνέχεια, ο προτεινόμενος αλγόριθμος ανακατασκευής εκτιμά το αραιό σήμα επιλύοντας επαναληπτικά ένα πρόβλημα βελτιστοποίησης με περιορισμούς χρησιμοποιώντας τη *Θεωρία Δυισμού* (*Duality Theory*) και τη μέθοδο των *υπο-κλίσεων* (*subgradients*). Εντούτοις, καθώς η οικογένεια των *Άλφα-Ευσταθών* κατανομών στερείται πεπερασμένης διασποράς, εισάγουμε μία τροποποιημένη *Lagrangian* συνάρτηση η οποία λαμβάνει υπόψη τη μη-Γκαουσιανή συμπεριφορά όπως εκφράζεται από τις λεγόμενες *Κλασματικές Ροπές Χαμηλότερης Τάξης* (*Fractional Lower-Order Moments*). Επιπλέον, αποδεικνύεται ότι η αντικειμενική συνάρτηση και οι περιορισμοί είναι διαχωρίσιμοι και επομένως αυτός ο αλγόριθμος επιδέχεται μια κατανεμημένη υλοποίηση από τους κόμβους ενός δικτύου αισθητήρων.

Η αυξημένη αποδοτικότητα των προτεινόμενων αλγορίθμων ανακατασκευής αραιού σήματος με χρήση συμπιεστικής δειγματοληψίας, με βάση μοντέλα που χαρακτηρίζονται από βαριές ουρές, επιδεικνύεται μέσα από συγκρίσεις με τις αποδόσεις μερικών εκ των κορυφαίων πρόσφατων τεχνικών, εφαρμόζοντάς τις σε μια σειρά πειραμάτων, από την ανακατασκευή προσομοιωμένων σημάτων και την επίλυση του προβλήματος εκτίμησης της Γωνίας Άφιξης (Direction-of-Arrival (DOA)), ως την ανάκτηση πραγματικών εικόνων από τους συντελεστές του μετασχηματισμού τους.

Contents

List of Figures **xix**

List of Tables **xxiii**

List of Abbreviations **xxiv**

List of Symbols **xxvi**

Structure of the thesis **xxix**

1 Introduction **1**
1.1 From Nyquist's sampling theorem to Compressed Sensing 2
1.1.1 Sensing/Encoding Part 4
1.1.2 Reconstruction/Decoding Part 7
1.2 Reconstruction techniques for CS 9
1.2.1 Deterministic (norm-based) reconstruction techniques 9
1.2.2 Probabilistic (Bayesian) reconstruction techniques 10
1.3 Distributed CS 10
1.4 Applications of CS 12
1.5 Contributions of the thesis 13

I Mathematical Background **15**

2 Wavelet Transform-a sparsity enforcing domain **17**
2.1 Introduction 17
2.2 Discrete Fourier Transform (DFT) 18
2.2.1 Short-Time Fourier Transform (STFT) 18
2.3 Wavelet Transform 19
2.3.1 The 1-Dimensional Wavelet Transform 19
2.3.2 The 2-Dimensional Discrete Wavelet Transform 23
2.4 Statistical properties of the DWT coefficients 25

3 Alpha-Stable models **27**
3.1 Introduction 27
3.2 The family of Stable Distributions 28
3.2.1 Univariate $S\alpha S$ Distributions 28
3.2.2 Multivariate Stable Distributions 31

3.2.3 Covariations . . . 33
3.3 Statistical Modeling of Wavelet Subband Coefficients . . . 36

II Heavy-tailed Bayesian Compressed Sensing Algorithms 39

4 Bayesian Compressed Sensing via Gaussian Scale Mixtures 41
4.1 Introduction . . . 41
4.2 Bayesian CS inversion . . . 42
4.3 Estimation of a sparse vector $\vec{w}$ using a GSM . . . 43
4.3.1 GSM prior model . . . 44
4.3.2 BCS inversion using a GSM . . . 44
4.4 Adaptive BCS using a GSM . . . 46
4.5 Performance evaluation: 1-D synthetic signals . . . 48
4.6 Performance evaluation: Image reconstruction . . . 57
4.7 Multiple-measurement BCS using a GSM . . . 65
4.7.1 Multiple-measurement vectors model . . . 66
4.7.2 BCS-GSM inversion using multiple measurement vectors . . . 66
4.7.3 Estimation of a sparse vector $\vec{w}$ . . . 68
4.7.4 Adaptive BCS-GSM using multiple measurement vectors . . . 70
4.8 Distributed implementation of MMV BCS-GSM . . . 71
4.8.1 Distributed MMV BCS-GSM using a star-shaped topology . . . 71
4.8.2 Distributed MMV BCS-GSM using a ring-shaped topology . . . 71
4.9 Performance evaluation: DOA estimation . . . 72
4.10 Performance evaluation: Image reconstruction . . . 79
4.11 Conclusions and future work . . . 85

5 BCS of Highly Impulsive Signals in Heavy-Tailed Noise 87
5.1 Introduction . . . 87
5.2 Statistical signal model . . . 88
5.3 Estimation of a sparse vector $\vec{w}$ via a multivariate Cauchy prior . . . 89
5.3.1 MMSE and MAP estimate of $\vec{w}$ . . . 90
5.3.2 Incremental basis selection via a tree-structure . . . 91
5.4 Performance evaluation . . . 94
5.5 Conclusions and future work . . . 97

6 Compressed Sensing using $S\alpha S$ Distributions 99
6.1 Introduction . . . 99
6.2 Statistical signal model . . . 100
6.2.1 A new $S\alpha S$ measurement matrix . . . 103
6.3 $S\alpha S$ minimum dispersion CS inversion . . . 103
6.3.1 Basis selection rule . . . 105
6.3.2 Estimation of p parameter . . . 106
6.4 Distributed $S\alpha S$-CS . . . 107
6.4.1 Dualization and distributed solution of the primal problem . . . 108
6.5 Performance evaluation . . . 113
6.5.1 Performance evaluation of $S\alpha S$-CS . . . 114
6.5.2 Performance evaluation of distributed $S\alpha S$-CS . . . 120
6.6 Conclusions and future work . . . 121

7 Conclusions and Future work **123**
7.1 Thesis overview 123
7.2 Future work 125

List of Figures

1.1 Fundamental blocks of a system described by the Nyquist's theorem. 2
1.2 Fundamental blocks of a CS-based system. 3
1.3 The CS measurement mappings. 5

2.1 The compression/encoding module of a CS-based system. 17
2.2 Comparison of time-frequency tiling for STFT and wavelet spectrum. 19
2.3 The filtering process of the 1-D DWT. 23
2.4 A 3-level 2-D DWT. 24
2.5 Original image "Facets", along with its 3-level 2-D DWT using Daubechies' 4 wavelet. 25

3.1 The decompression/reconstruction module of a CS-based system. 27
3.2 Standard $S\alpha S$ density functions for different values of the characteristic exponent α (parameterization (3.2)). 29
3.3 Tails of the standard $S\alpha S$ density functions for different values of the characteristic exponent α (parameterization (3.2)). 30
3.4 Simulated $S\alpha S$ sequences for different values of the characteristic exponent. . . 31
3.5 Curves representing the standard deviation of the $\hat{\mathbf{c}}^{FLOM}$ covariation estimator as a function of q for $\alpha = 1.2$, 1.5 and 25 dispersion pairs (γ_X, γ_Y). 36
3.6 Acoustic signal and APD curves corresponding to the empirical, Gaussian, GGD and $S\alpha S$ distributions, at each decomposition level (3-level DWT (db4)). 38
3.7 "Indor 4" image and APD curves corresponding to the empirical, Gaussian, GGD and $S\alpha S$ distributions, at each decomposition level (3-level DWT (db4)). 38

4.1 Reconstruction of uniform spikes for $N = 512$, $L = 20$, $M = 110$. (a) Original signal, (b) Reconstruction with BCS, $\|\vec{x} - \vec{x}_{BCS}\|_2/\|\vec{x}\|_2 = 0.0114$, (c) Reconstruction with BCS-GSM, $\|\vec{x} - \vec{x}_{BCS-GSM}\|_2/\|\vec{x}\|_2 = 0.0106$. 48
4.2 Reconstruction of non-uniform spikes for $N = 512$, $L = 20$, $M = 110$. (a) Original signal, (b) Reconstruction with BCS, $\|\vec{x} - \vec{x}_{BCS}\|_2/\|\vec{x}\|_2 = 0.0122$, (c) Reconstruction with BCS-GSM, $\|\vec{x} - \vec{x}_{BCS-GSM}\|_2/\|\vec{x}\|_2 = 0.0121$. 49
4.3 Reconstruction performance of BCS-GSM and BCS: (a) Average reconstruction error, and (b) Average number of non-zero weights, over 100 Monte-Carlo runs using random signals with 20 non-uniform spikes ($\sigma_\eta = 0.005$). 50
4.4 Average CS ratio as a function of M for the BCS-GSM, BCS methods ($L = 20$, $\sigma_\eta = 0.005$). 51
4.5 Average reconstruction error as a function of CS ratio for the BCS-GSM, BCS methods ($L = 20$, $\sigma_\eta = 0.005$). 51

4.6 Average RREs and CS ratios as a function of SNR for the BCS-GSM and the selected CS methods for sparse signals with uniform spikes. 54
4.7 Average RREs and CS ratios as a function of SNR for the BCS-GSM and the selected CS methods for sparse signals with non-uniform spikes. 55
4.8 Average reconstruction error of the standard (non-adaptive) and adaptive BCS-GSM, over 100 Monte-Carlo runs using random signals with 20 spikes ($\sigma_\eta = 0.005$). 56
4.9 Average RREs and CS ratios over 100 Monte-Carlo runs for the adaptive BCS-GSM and BCS methods using random signals with $L = 15$ uniform spikes ($N = 512$, $\sigma_\eta = 0.01$). 56
4.10 Test images. 57
4.11 PSNRs and CS ratios for "*Indor 2*" image as a function of the number of CS measurements for SNR = 5, 10 dB. 58
4.12 PSNR and CS ratios of the BCS-GSM, BCS and their adaptive implementations for "*Indor 2*" image. 58
4.13 Medical images (128×128) used for evaluation of the performance of BCS-GSM. 59
4.14 PSNRs for CS reconstruction of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of the number of measurements and for SNR $= 7.5, 15$ dB. 61
4.15 PSNRs for CS reconstruction of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of the number of measurements and for SNR $= 7.5, 15$ dB. 61
4.16 CS ratios for CS reconstruction of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of the number of measurements and for SNR $= 7.5, 15$ dB. 62
4.17 CS ratios for CS reconstruction of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of the number of measurements and for SNR $= 7.5, 15$ dB. 62
4.18 Original and CS reconstructed images of "Aneu3" for SNR $= 7.5, 15$ dB. 63
4.19 Original and CS reconstructed images of "Cisternogram" for SNR $= 7.5, 15$ dB. 63
4.20 Original and CS reconstructed images of "CoronaryAngio" for SNR $= 7.5, 15$ dB. 64
4.21 Original and CS reconstructed images of "CTMyeloL1" for SNR $= 7.5, 15$ dB. . 64
4.22 Star-shaped and ring-shaped topologies of a sensor network implementing the distributed MMV BCS-GSM method. 68
4.23 DOA estimation performance for one source (54^o). 74
4.24 DOA estimation performance for two sources ($41^o, 44^o$). 75
4.25 Average percentages of successful DOA estimations for a varying number of sources and sensors. 77
4.26 Average CS ratios for a varying number of sources and sensors. 77
4.27 Average percentages of successful DOA estimations for a varying number of sources and initial SNR values at the leftmost sensor ($K = 5$). 78
4.28 Average CS ratios for a varying number of sources and initial SNR values at the leftmost sensor ($K = 5$). 78
4.29 PSNRs for CS reconstruction using MMV BCS-GSM of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB). 80
4.30 PSNRs for CS reconstruction using MMV BCS-GSM of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB). 80
4.31 CS ratios for CS reconstruction using MMV BCS-GSM of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB). 81
4.32 CS ratios for CS reconstruction using MMV BCS-GSM of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB). 81

4.33 PSNRs comparison between Linear (optimal) reconstruction with BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M (SNR $= 7.5, 15$ dB). 82
4.34 PSNRs comparison between Linear (optimal) reconstruction with BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M (SNR $= 7.5, 15$ dB). 82
4.35 CS ratios comparison between BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M (SNR $= 7.5, 15$ dB). 83
4.36 CS ratios comparison between BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M (SNR $= 7.5, 15$ dB). 83
4.37 Original and noisy (SNR = 5 dB) CS reconstructed images of "Indor 2" using Linear, BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) methods. 84
4.38 Original and noisy (SNR = 5 dB) CS reconstructed images for "Indor 4" using Linear, BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) methods. 84
4.39 Original and noisy (SNR = 5 dB) CS reconstructed images for "Nemasup" using Linear, BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) methods. 85

5.1 Simulated standard $S\alpha S$ sequences along with their detail wavelet coefficients histograms (3 levels, "db 4"). 88
5.2 Average MMSE and MAP reconstruction errors for FBMP and MvC methods as a function of M (SNR=10, 15 dB). 95
5.3 Sparsity performance of FBMP and MvC methods. 96
5.4 SpR ratio for FBMP and MvC methods as a function of M (SNR=10, 15 dB). 97
5.5 Average MMSE and MAP reconstruction errors for FBMP and MvC methods as a function of α (SNR=8, 10 dB). 98
5.6 SpR ratio for FBMP and MvC methods as a function of α (SNR=8, 10 dB). . . 98

6.1 Test stability property for four kinds of measurement matrices $\mathbf{\Phi}$ (ref. in the text).103
6.2 Network topology for implementing the distributed $S\alpha S$-CS algorithm. 112
6.3 Noise dispersion contours: (a) Noise dispersion contours in the jointly $S\alpha S$ case, as a function of γ_g and FSNR, for $\alpha_g = 1.3, 1.45, 1.7, 1.95$, (b) Noise dispersion contours as a function of γ_g and α_η, for $\alpha_g = 1.3, 1.45, 1.7, 1.95$ and FSNR $= 10$ dB.115
6.4 Average rSNR's as a function of the number of CS measurements M ($\alpha = 1.3$, $\gamma_w = 0.7$ and FSNR $= 10$ dB). 116
6.5 Average CS ratios of the CS-based recovery methods as a function of the number of CS measurements M ($\alpha = 1.3$, $\gamma_w = 0.7$ and FSNR $= 10$ dB). 116
6.6 Average rSNR's as a function of the number of CS measurements M ($\alpha = 1.5$, $\gamma_w = 1$ and FSNR $= 10$ dB). 117
6.7 Average CS ratios of the CS-based recovery methods as a function of the number of CS measurements M ($\alpha = 1.5$, $\gamma_w = 1$ and FSNR $= 10$ dB). 117
6.8 Average rSNR for $S\alpha S$-CS and LASSO as a function of the number of CS measurements M, for $\gamma_w = 1$, FSNR $= 10$ dB and $\alpha = 1.4, 1.5, 1.7, 1.8$. 118
6.9 Average CS ratios for $S\alpha S$-CS and LASSO as a function of the number of CS measurements M, for $\gamma_w = 1$, FSNR $= 10$ dB and $\alpha = 1.4, 1.5, 1.7, 1.8$. 118
6.10 (a) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and FSNR (in dB) for $\gamma_w = 1$ and $M = 100$, (b) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and FSNR (in dB) [Top view]. 119
6.11 Average CS ratio for $S\alpha S$-CS and LASSO as a function of α and FSNR (in dB) for $\gamma_w = 1$ and $M = 100$. 119

6.12 (a) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and α_η for $\gamma_w = 1$, $M = 100$ and FSNR = 8 dB, (b) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and α_η [Top view]. 120
6.13 Average CS ratio for $S\alpha S$-CS and LASSO as a function of α and α_η for $\gamma_w = 1$, $M = 100$ and FSNR = 8 dB. 120
6.14 Average percentage of successful retrievals of the significant basis functions for the standard and FLOM-based Lagrangian function, as a function of α and FSNR (in dB). 122
6.15 (a) Average reconstruction rSNR (in dB) for the "centralized" $S\alpha S$-CS method, as a function of α and FSNR (in dB), (b) Average reconstruction rSNR (in dB) for the distributed $S\alpha S$-CS method, as a function of α and FSNR (in dB). . . . 122

List of Tables

3.1 Optimal q parameter as a function of the characteristic exponent α. 36

4.1 Performance comparison in terms of PSNR and CS ratio values for the reconstruction of "*Indor 4*" and "*Nemasup*" images (noise-free & noisy versions) with $c = 0.6$ ($M = 2611$). 59

List of Abbreviations

APD	**A**mplitude **P**robability **D**ensity
BCS	**B**ayesian **C**ompressed **S**ensing
BCS-GSM	**B**ayesian **C**ompressed **S**ensing using a **G**aussian **S**cale **M**ixture
BP	**B**asis **P**ursuit
CI	**C**ompressive **I**maging
CS	**C**ompressed **S**ensing
CWT	**C**ontinuous **W**avelet **T**ransform
DCS	**D**istributed **C**ompressed **S**ensing
DEC	**D**ifferential **E**ntropic **C**lustering
DFT	**D**iscrete **F**ourier **T**ransform
DOA	**D**irection **o**f **A**rrival
DWT	**D**iscrete **W**avelet **T**ransform
FLOM	**F**ractional **L**ower-**O**rder **M**oment
GGD	**G**eneralized **G**aussian **D**istribution
GPSR	**G**radient **P**rojection for **S**parse **R**econstruction
GSM	**G**aussian **S**cale **M**ixture
i.i.d.	**i**ndependent **i**dentically **d**istributed
L1EQ-PD	ℓ_1-norm minimization using the Primal-Dual interior point method
MAP	**M**aximum **a** **P**osteriori
ML	**M**aximum **L**ikelihood
MMV	**M**ultiple **M**easurement **V**ector
MRI	**M**agnetic **R**esonance **I**maging
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
RIP	**R**estricted **I**sometry **P**roperty
SαS	**S**ymmetric Alpha-**S**table
SMV	**S**ingle **M**easurement **V**ector
SNR	**S**ignal-to-**N**oise **R**atio
SOCP	**S**econd-**O**rder **C**one **P**rogram
STFT	**S**hort-**T**ime **F**ourier **T**ransform
StOMP	**St**agewise **O**rthogonal **M**atching **P**ursuit
UUP	**U**niform **U**ncertainty **P**rinciple

List of Symbols

$\mathbf{A}$	matrix
$\mathbf{A}^{-1}$	inverse matrix
$\mathbf{A}^{T}$	transpose matrix
$\mathbf{A}^{H}$	Hermittian matrix
$\lvert\mathbf{A}\rvert$	matrix determinant
$\vec{x}$	column vector (unless stated otherwise to be a row vector)
$\vec{x}^{T}$	transpose (row) vector
$\mathcal{O}(\cdot)$	order of $(\cdot)$
$\lVert\cdot\rVert$	vector/matrix norm
log	natural logarithm
$P\{\mathcal{A}\}$	probability of event $\mathcal{A}$
$\mathbb{E}\{\cdot\}$	Expectation operator
$Var\{X\}$	Variance of random variable X
$\lvert a\rvert$	absolute value ($a \in \mathbb{R}$)
$abs(\vec{x})$	$[\lvert x_1\rvert, \ldots, \lvert x_N\rvert]^T$, $\vec{x} \in \mathbb{R}^N$
α-SG($\mathbf{R}$)	sub-Gaussian $S\alpha S$ vector with underlying covariance matrix $\mathbf{R}$
$z^{<a>}$	$\lvert z\rvert^{a-1}\bar{z}$, $z \in \mathbb{C}$, $a \geq 0$
$\bar{z}$	complex conjugate
$sign(\cdot)$	signum function
$\mathcal{S}_\alpha$	linear space of jointly $S\alpha S$ random variables
$[X, Y]_\alpha$	covariation between two jointly $S\alpha S$ random variables
λ_{XY}	covariation coefficient of X with Y
$Corr_\alpha(X, Y)$	symmetric covariation coefficient of X with Y
$\lVert X\rVert_\alpha$	covariation norm of $X \in \mathcal{S}_\alpha$

Structure of the thesis

The present thesis deals with the problem of reconstructing a given signal which is sparse by itself, or it can be sparsified in an appropriate transform domain, by employing a vector containing much less measurements than the original signal size. In particular, since we work in a statistical framework, the prior knowledge that the signal is sparse is exploited by modelling the coefficients vector via members of families including distributions with heavy (algebraic) tails and specifically, of the family of *Alpha-Stable* distributions. Then, efficient Bayesian and iterative methods are proposed exploiting the accuracy of the *Alpha-Stable* family in capturing the sparse nature of (highly) impulsive signals. Thus, the problem of reconstructing a signal from its associated measurement vector is reduced to a problem of estimating the corresponding model parameters. The thesis is organized as follows:

Chapter 1
This chapter provides the necessary background of the recently introduced theory of *Compressed Sensing* (CS). We discriminate between its two structural components, namely, the encoding/sensing and the decoding/reconstruction part and discuss their specific inherent requirements. Besides, we classify most of the existing state-of-the-art CS reconstruction algorithms according to whether they proceed in a deterministic or in a probabilistic (Bayesian) framework. This makes the comparison with the proposed approaches more meaningful and convenient.

Chapter 2
This chapter illustrates the fact that many natural signals, which may be not sparse by themselves, can be sparsified in the wavelet transform domain. It also sets out the main concepts and properties concerning the wavelet analysis, which is employed in some of the subsequent experimental evaluations.

Chapter 3
In this chapter we review the basic theory regarding the family of Alpha-Stable models, which will be the fundamental statistical tool for developing the proposed CS reconstruction algorithms.

Chapter 4
This chapter describes the first of the proposed CS reconstruction algorithms. In particular, the fact that the original signal, or a transformed version of it, is sparse (or compressible) is exploited by modeling the statistics of the coefficients vector with a Gaussian Scale Mixture

(GSM), which is suitable in approximating a heavy-tailed behavior. The problem of signal reconstruction is reduced to the problem of estimating the GSM model parameters, which is solved using a fast iterative approach. Then, this method is extended in the case when multiple observations of a single sparse signals are available. For instance, this is the case of a sensor network application where multiple sensors acquire the same signal of interest. The proposed multiple-observation CS method also exploits the property, that due to the acquisition process, the multiple observation vectors will be characterized by a common sparsity structure with high probability. The proposed methods are then compared with recent state-of-the-art techniques for reconstructing simulated signal, as well as in the problem of Direction-of-Arrival (DOA) estimation and in the recovery of real-world images from their wavelet transform coefficients.

Chapter 5

In this chapter, the second proposed method is presented for solving the problem of reconstructing a highly impulsive signal corrupted by heavy-tailed noise. The prior belief that the vector of coefficients should be highly sparse is enforced by fitting its prior distribution by means of a (heavy-tailed) multivariate Cauchy distribution, which is a member of the sub-Gaussian family. In addition, a multivariate Cauchy distribution is also employed to model the heavy-tailed behavior of the noise corrupting the coefficients, since it models efficiently highly impulsive environments and also it yields closed form expressions in the subsequent Bayesian inference. The experimental results show that our proposed method achieves an improved reconstruction performance, in terms of a smaller reconstruction error, while increasing the sparsity using less basis functions, when compared with a recently introduced Gaussian-based Bayesian implementation and with previous norm-based CS reconstruction algorithms.

Chapter 6

In this chapter, the two previous methods are generalized by developing a CS method which employees an arbitrary *symmetric Alpha-Stable* distribution and thus, it provides additional degrees of freedom to the considered statistical model. The proposed CS algorithm for estimating a sparse vector proceeds by solving iteratively a constrained optimization problem using the duality theory and the method of subgradients. However, since the family of *symmetric Alpha-Stable* distributions lacks finite variance, we introduce a modified Lagrangian function which takes into account the true non-Gaussian behavior as expressed by the so called *Fractional Lower-Order Moments*. In addition, it is shown that the objective function and the constraints are separable and thus, this algorithm is amenable to a distributed implementation from the nodes of a sensor network. The experimental results reveal an increased reconstruction performance in terms of a reduced reconstruction error, while achieving an increased sparsity, when compared with state-of-the-art iterative CS methods.

Chapter 7

This chapter serves as a conclusion and summarization of the main results of this thesis and provides directions for future work.

Chapter

1 Introduction

The whole is more than the sum of the parts.

Aristotle (384 B.C.-322 B.C.)
Metaphysica

During the last decades, information is being gathered and processed at an explosive rate. This fact gives rise to a very important issue, that is, how to effectively and precisely describe the information content of a given source signal or an ensemble of source signals, such that it can be stored, processed or transmitted by taking into consideration the limitations and capabilities of the several digital devices.

One of the fundamental principles of signal processing for decades is the *Nyquist-Shannon sampling theorem* [1, 2]:

Theorem 1.1 (Nyquist-Shannon sampling theorem) *If a function $x(t)$ contains no frequencies higher than B cps (cycles per second), it is completely determined by giving its ordinates at a series of points spaced $\frac{1}{2B}$ seconds apart.*

In essence this theorem states that the number of samples needed to reconstruct a signal without error is dictated by its bandwidth - the length of the shortest interval which contains the support of the spectrum of the signal. In particular, the signal can be perfectly reconstructed from its samples if the sampling rate exceeds $2B$ samples per second, where B is the highest frequency in the original signal. The reconstruction is simply a linear interpolation with a "sinc" kernel. If a signal contains a component at exactly B cps (Hertz), then samples spaced at exactly $\frac{1}{2B}$ seconds do not completely determine the signal, Shannon's statement notwithstanding.

The above theorem actually describes two basic processes in signal processing: a *sampling process (encoding)*, in which a continuous-time signal is converted to a discrete-time signal and a *reconstruction process (decoding)*, in which the original continuous signal is recovered from the discrete-time signal. Figure 1.1 on the following page shows the fundamental blocks of a system as described by the Nyquist's theorem. A few consequences that can be drawn are the following:

1. If the highest frequency B in the original signal is known, the theorem gives the lower bound on the sampling frequency for which perfect reconstruction can be assured. This lower bound to the sampling frequency, $2B$, is called the Nyquist rate.

2. If instead the sampling frequency is known, the theorem gives an upper bound for the frequency components of the signal to allow for perfect reconstruction. This upper bound is the Nyquist frequency, denoted by B_N.

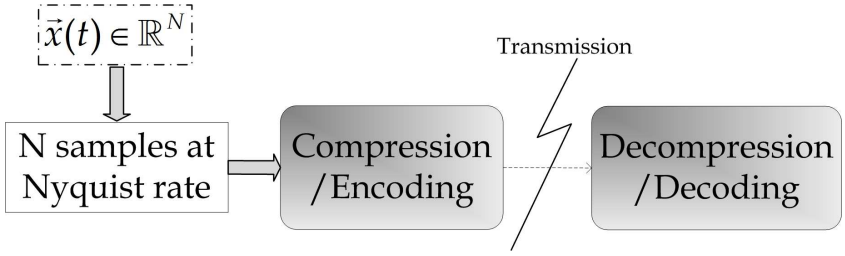


Figure 1.1: Fundamental blocks of a system described by the Nyquist's theorem.

Both of these cases imply that the signal to be sampled must be *bandlimited*, that is, any component of this signal which has a frequency above a certain threshold should be zero, or at least sufficiently close to zero to allow us to neglect its influence on the resulting reconstruction. In practice, neither of the two statements of the sampling theorem described above can be completely satisfied and neither can the reconstruction formula be precisely implemented. Furthermore, in practice, a signal can never be perfectly bandlimited, since ideal "brick-wall" filters cannot be realized. All practical filters can only attenuate frequencies outside a certain range and cannot remove them entirely. In addition, a "time-limited" signal can never be bandlimited. This means that even if an ideal reconstruction could be made, the reconstructed signal would not be exactly the original signal (the induced error is referred to as aliasing, meaning that the same set of samples may describe many different signals).

The Nyquist-Shannon sampling theorem is also known to be a *sufficient condition*. In fact, this underlies nearly all signal acquisition tasks, such as in audio and visual electronics, medical imaging devices and radio receivers. However, there are many cases in our everyday life in which sampling at the Nyquist rate results in too many data and thus, demanding an increased processing power, as well as storage requirements. For instance, assume the acquisition of an image[1] using a common digital camera of 5.0 Megapixels. With uniform sampling at the standard Nyquist rate we obtain approximately $N = 5{\cdot}10^6$ samples. By using the RGB standard (3 colors per sample) with 8 bits/color, we get a digital image of size equal to $5 \cdot 10^6 \times 3 \times 8 = 120 \cdot 10^6$ million bits, which is a large amount of data. Thus, the use of an appropriate compression scheme is necessary to reduce the size of the dataset.

1.1 From Nyquist's sampling theorem to Compressed Sensing

The essential purpose of sensing and sampling systems is to capture accurately the salient information in a signal of interest. As mentioned before the bandlimited behavior of a signal, when compared with the assumptions of the Nyquist's sampling theorem, is sufficient to ensure that the signal reconstructed from a set of uniform samples is unique and equal to the original signal. In a standard data acquisition process massive amounts of data are collected only to be - in large part - discarded at the compression stage to facilitate storage and transmission, which is extremely wasteful. For instance, in the case of the digital camera described before, the sensing equipment consists of millions of imaging sensors, the pixels, but eventually encodes the picture in just a few hundred kilobytes.

A mathematical theory that emerged recently, presents the background for developing a novel sensing/sampling paradigm that goes against the common tenet in data acquisition. As

[1]For signals, such as images that are not naturally bandlimited, the sampling rate is dictated not by the Nyquist's theorem but by the desired spatial and/or temporal resolution. However, a common approach in such systems is to use an anti-aliasing low-pass filter to bandlimit the signal before sampling, and so the Nyquist's theorem appears implicitly.

modern electronic equipment acquires and exploits ever-increasing amounts of data, it has already become a common perception that most of the data we acquire can be discarded with almost no perceptual loss - confirmed by the broad success of lossy compression formats for audio, images and other specialized data. The phenomenon of ubiquitous *compressibility* raises very natural questions: "Why should we give so much effort to acquire all the data when *most* of what we get will be discarded?", "Can't we just *directly measure* the part of information that will not end up being thrown away?". These questions lead to a more fundamental theoretical problem: "Could the processes of sampling and compression/encoding be combined into a single process?". The answer to this question is given by theory of *compressed sensing*.

Compressed sensing (CS), also known as compressive sensing, compressive sampling and sparse sampling, is a technique for acquiring and reconstructing a signal utilizing the prior knowledge that it is *sparse* or *compressible*, which provides a stricter sampling condition yielding a sub-Nyquist sampling criterion. The field has existed for at least four decades, but recently the field has exploded, in part due to several important works [3, 4, 5]. CS relies on two basic principles, namely, the *sparsity*, which refers to the signals of interest, and the *incoherence*, which refers to the sensing modality.

Sparsity expresses the fact that the "information rate" of a continuous-time signal may be much smaller than suggested by its bandwidth via the Nyquist's theorem, or that a discrete-time signal depends on a number of degrees of freedom which is comparably much smaller than its (finite) length. More precisely, CS exploits the property that many natural signals are sparse or compressible in the sense that they result in a highly compact representation when they are projected on an appropriate set of localized orthonormal basis functions $\mathbf{\Psi}$ (e.g.,wavelets and sinusoids), as several works [6, 7, 8] have shown. On the other hand, incoherence extends the duality between the domains of time and frequency and expresses the fact that objects (signals, images) having a sparse representation in $\mathbf{\Psi}$ must be spread out in the domain in which they are acquired, just as a Dirac or a spike in the time-domain is spread out in the frequency-domain.

Figure 1.2 presents the corresponding fundamental parts of a CS-based sensing system, analogous to the one shown in Figure 1.1. Notice that when working in the CS framework, the processes of sensing and compression/encoding are merged into a single process. The performance of a CS-based acquisition system is determined independently by its two main structural components, namely, the *sensing/encoding* part and the *reconstruction/decoding* part. In the following, we review the main factors that affect the behavior of each component separately. A more detailed formation will be presented in the next chapter.

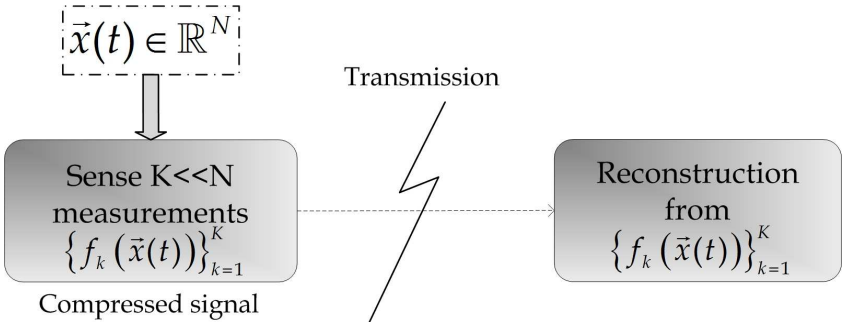


Figure 1.2: Fundamental blocks of a CS-based system.

1.1.1 Sensing/Encoding Part

The performance of the first component, namely the sensing or encoding module, is in direct dependence on the degree of sparsity of a (discrete-time)[2] signal $\vec{x} \in \mathbb{R}^N$. Let $\mathbf{\Psi} = [\vec{\psi}_1, \vec{\psi}_2, \ldots, \vec{\psi}_N]$ be a $N \times N$ matrix, whose columns correspond to the frequency-domain (or more generally to a transform-domain) basis functions. Then, signal $\vec{x}$ can be represented as

$$\vec{x} = \mathbf{\Psi}\vec{w} \ , \tag{1.1}$$

where $\vec{w} \in \mathbb{R}^N$ is the *weight vector*. Obviously, $\vec{x}$ and $\vec{w}$ are equivalent representations of the signal, with $\vec{x}$ being in the time (or space) domain and $\vec{w}$ in the (transform) $\mathbf{\Psi}$ domain. Then, $\vec{x}$ is called L-sparse in basis $\mathbf{\Psi}$ ($L \ll N$) if $\vec{w}$ has only L non-zero components. Notice that, in general, we will consider sparsity in an appropriate transform domain. However, the case of $\mathbf{\Psi} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix, is reduced to considering that the original signal $\vec{x} \equiv \vec{w}$ is sparse by itself.

However, note that this strict definition of sparsity is very restrictive and does not include many real-world signals. A more precise way to define sparsity would be to say that $\vec{x}$ is L-sparse if it is "well-approximated" by a linear combination of L basis vectors from $\mathbf{\Psi}$. In this case we say that $\vec{x}$ is L-compressible in basis $\mathbf{\Psi}$ when the sorted transform coefficients follow a power-decay law:

$$|w_{(l)}| \lesssim \mathcal{O}(l^{-\alpha}) \ , \ \alpha > 1 \ , \tag{1.2}$$

where $w_{(l)}$ is the l-th largest coefficient, resulting in the following error for the best L-term approximation:

$$\|\vec{w} - \vec{w}_L\|_2 \lesssim L^{-\alpha+\frac{1}{2}} \ , \tag{1.3}$$

where $\vec{w}_L$ denotes the approximation of $\vec{w}$ using the L largest (in absolute value) coefficients - with the remaining ones set to zero - and $\|\vec{v}\|_2$ is the ℓ_2-norm of a vector $\vec{v} \in \mathbb{R}^N$ given by $\|\vec{v}\|_2 \triangleq \left(\sum_{i=1}^N |v_i|^2\right)^{1/2}$.

Consider also an $M \times N$ measurement matrix $\mathbf{\Phi} = [\vec{\phi}_1, \vec{\phi}_2, \ldots, \vec{\phi}_M]^T$, $M < N$, where the rows of $\mathbf{\Phi}$ are incoherent with the columns of $\mathbf{\Psi}$ (the incoherence assumption mentioned before). In simple words this means that the rows of $\mathbf{\Phi}$ cannot sparsely represent the columns of $\mathbf{\Psi}$ and vice versa. For example, let $\mathbf{\Phi}$ contain independent and identically distributed (i.i.d.) Gaussian entries. Such a matrix is incoherent with any fixed transform matrix $\mathbf{\Psi}$ with high probability (universality property) [4].

If the signal $\vec{x}$ is compressible in $\mathbf{\Psi}$, then, it is possible to perform directly a compressed set of measurements $\vec{g}$, resulting in a simplified sensing system. The relation between the original signal $\vec{x}$ and the compressed sensing (CS) measurements $\vec{g}$ is obtained through random projections. That is, the m-th CS measurement, $\vec{g}_m$, results by projecting $\vec{x}$ onto a random linear combination of the basis functions, $\vec{g}_m = \vec{x}^T(\mathbf{\Psi}\vec{\phi}_m)$, where $\vec{\phi}_m \in \mathbb{R}^N$ is a random vector with i.i.d. components. The CS measurements can be written in the following compact form,

$$\vec{g} = \mathbf{\Phi}\mathbf{\Psi}^T\vec{x} \overset{(1.1)}{=} \mathbf{\Phi}\vec{w} \ . \tag{1.4}$$

Figure 1.3 shows a graphical representation of Eq. (1.4). A problem could arise if two distinct L-sparse signals, $\vec{x}$ and $\vec{x}'$, are mapped to the same compressed data, that is, $\vec{g} = \vec{g}'$. This is exactly the situation where certain sparse vectors lie in the null space of the measurement matrix $\mathbf{\Phi}$. Matrices that are resilient to this ambiguity are those that satisfy the so-called Restricted Isometry Property (RIP) (sometimes also called the Uniform Uncertainty Principle (UUP)) [5].

[2] In the subsequent analysis we ignore the time dependence for convenience and write $\vec{x}$ instead of $\vec{x}(t)$.

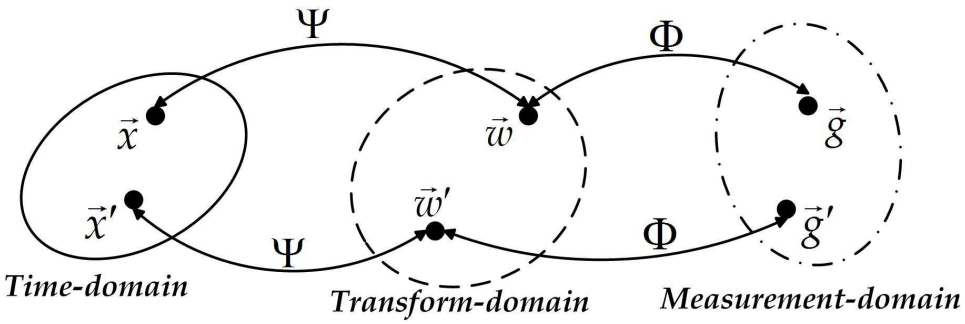


Figure 1.3: The CS measurement mappings.

Definition 1.1 (RIP) *A $M \times N$ measurement matrix $\mathbf{\Phi}$ with unit-norm rows is said to satisfy the RIP of order L whenever*

$$(1-\epsilon_L)\frac{M}{N}\|\vec{w}\|_2^2 \leq \|\mathbf{\Phi}\vec{w}\|_2^2 \leq (1+\epsilon_L)\frac{M}{N}\|\vec{w}\|_2^2 \tag{1.5}$$

holds simultaneously for all L-sparse vectors $\vec{w} \in \mathbb{R}^N$ for sufficiently small values of ϵ_L.

To clarify the connection between the RIP and CS, suppose we acquire L-sparse signals via the measurement matrix $\mathbf{\Phi}$. Assume also that $\epsilon_{2L} \ll 1$. This implies that all pairwise distances between L-sparse signals must be well preserved in the measurement domain. That is,

$$(1-\epsilon_{2L})\|\vec{w}_1 - \vec{w}_2\|_2^2 \leq \|\mathbf{\Phi}\vec{w}_1 - \mathbf{\Phi}\vec{w}_2\|_2^2 \leq (1+\epsilon_{2L})\|\vec{w}_1 - \vec{w}_2\|_2^2 \ , \tag{1.6}$$

holds for all L-sparse vectors $\vec{w}_1$, $\vec{w}_2$. This condition guarantees the existence of efficient and robust algorithms for discriminating L-sparse signals based on their compressive measurements.

In practice, measurement matrices that satisfy the RIP are easy to generate. It has been established [4, 9] that $M \times N$ matrices $\mathbf{\Phi}$ whose entries are i.i.d. realizations of certain zero-mean random variables satisfy the RIP with very high probability when $M \geq c \cdot \log(N) \cdot L$, where c is a positive constant. Besides, physical limitations of real sensing systems motivate the unit-norm restriction on the rows of $\mathbf{\Phi}$, which essentially limits the amount of "sampling energy" allocated to each measurement. Some families of matrices that satisfy the RIP with very high probability are the following:

- ✔ *Gaussian matrices*: the entries of $\mathbf{\Phi}$ are independently sampled from a zero-mean Gaussian distribution with variance $1/M$. Then, the exact reconstruction of the L-sparse weight vector $\vec{w}$, and consequently the exact recovery of the original signal $\vec{x}$, can be achieved with probability $1-\mathcal{O}(e^{-\gamma N})$ (for some $\gamma > 0$) if sparsity, L, and the number of measurements, M, satisfy the following inequality:

$$L \leq c \cdot \frac{M}{\log\left(\frac{N}{M}\right)} \ . \tag{1.7}$$

- ✔ *Binary matrices*: the entries of $\mathbf{\Phi}$ are independently sampled from the symmetric Bernoulli distribution, that is, $\mathrm{P}\{[\mathbf{\Phi}]_{mn} = \pm 1/\sqrt{M}\} = 1/2$, where $\mathrm{P}\{A\}$ denotes the probability of event A and $[\mathbf{\Phi}]_{mn}$ is the entry of $\mathbf{\Phi}$ on the m-th row and the n-th column. Then, as in the Gaussian case, the exact recovery of $\vec{x}$ is guaranteed with probability $1-\mathcal{O}(e^{-\gamma N})$ ($\gamma > 0$) if inequality (1.7) is satisfied.

- ✔ *Fourier matrices*: the matrix $\mathbf{\Phi}$ is a partial Fourier matrix obtained by selecting M rows uniformly at random (random Fourier samples - sinusoids) and re-normalizing them so

that they have unit-norm. Then, in [5] it was shown that the exact recovery of $\vec{x}$ is guaranteed with probability close to 1 if the following inequality holds:

$$L \leq c \cdot \frac{M}{\big(\log(N)\big)^6} \ . \tag{1.8}$$

Recently [10], this upper bound was improved: $L \leq c \cdot M/\big(\log(N)\big)^4$.

✔ *Incoherent matrices*: as it was mentioned before, incoherence (together with sparsity) is a basic principle of the sensing modality. In addition, it generalizes the above three families of matrices that satisfy the RIP. The incoherence between the measurement matrix $\mathbf{\Phi}$ and the transform basis $\mathbf{\Psi}$ is measured by the following coefficient:

$$\mu(\mathbf{\Phi}, \mathbf{\Psi}) = \sqrt{N} \cdot \max_{1 \leq k,j \leq N} |\vec{\phi}_k^T \vec{\psi}_j| \ . \tag{1.9}$$

In particular, the smaller the value of $\mu(\mathbf{\Phi}, \mathbf{\Psi})$, the larger the incoherence between the two matrices. Select the columns of $\mathbf{\Phi}$ uniformly at random by orthonormalizing N vectors sampled independently and uniformly on the unit sphere in $\mathbb{R}^M$. Then, with high probability, the coherence between $\mathbf{\Phi}$ and $\mathbf{\Psi}$ is approximately $\sqrt{2\log(N)}$ for *any* fixed matrix $\mathbf{\Psi}$. In [11] it was shown that, when the pair $(\mathbf{\Phi}, \mathbf{\Psi})$ is highly incoherent, the exact recovery of $\vec{x}$ is guaranteed with probability that exceeds $1 - \vartheta$ if the following inequality holds:

$$L \leq c \cdot \frac{1}{\mu^2(\mathbf{\Phi}, \mathbf{\Psi})} \cdot \frac{M}{\log\big(\frac{N}{\vartheta}\big)} \ . \tag{1.10}$$

An example of such a pair of matrices consists of a matrix $\mathbf{\Phi}$ whose columns contain a spike basis ($\{\vec{\phi}_k(n) = \delta(n-k)\}$, $k = 1, \ldots, M$, $n = 1, \ldots, N$, where $\delta(\cdot)$ is the Dirac function) and a matrix $\mathbf{\Psi}$ whose columns form a Fourier basis ($\{\vec{\psi}_j(n) = N^{-1/2} e^{i2\pi jn/N}\}$, $j = 1, \ldots, N$, $n = 1, \ldots, N$).

Another example consists of a matrix $\mathbf{\Phi}$ whose columns contain a noiselet basis [12] and a matrix $\mathbf{\Psi}$ whose columns form a wavelet basis [6].

Inequality (1.10) reveals the prominent role of incoherence in the design of the sensing/encoding part of a CS-based sensing system: the larger the incoherence (smaller $\mu(\mathbf{\Phi}, \mathbf{\Psi})$), the fewer measurements (M) are needed for a fixed degree of sparsity (L). In particular, if $\mu(\mathbf{\Phi}, \mathbf{\Psi}) \simeq 1$ (maximal incoherence) then $\mathcal{O}(L \cdot \log(N))$ measurements suffice to reconstruct accurately the original signal, instead of acquiring N samples. Moreover, since the two fundamental components of a CS-based system, namely the sensing part and the reconstruction part, are completely decoupled, the construction of pairs $(\mathbf{\Phi}, \mathbf{\Psi})$ that promote an increased sparsity and a higher incoherence can improve the overall system's performance for a fixed reconstruction approach.

The most important is that a high incoherence is actually the only requirement of the sensing module. The signal $\vec{x}$ can be recovered exactly from the set of CS measurements $\vec{g}$ by solving specific optimization problems, described in the following section, which do not assume *any knowledge* about the sparsity of $\vec{x}$ in basis $\mathbf{\Psi}$, that is, about the number of non-zero components (L) of $\vec{w}$, their locations, or their amplitudes which are assumed to be completely unknown a priori. Thus, the CS theory suggests that it only suffices to sample non-adaptively and incoherently. Following this process would essentially acquire the signal directly in a *compressed form*. All that is needed is a decoder to reconstruct/decode this data, as described in the next section.

1.1.2 Reconstruction/Decoding Part

With the so far discussion, it is clear that the problem of reconstructing the original signal $\vec{x}$ from the CS measurements $\vec{g}$, is equivalent to estimating the L-sparse weight vector $\vec{w}$. In the subsequent analysis we will assume that the pair of matrices $(\mathbf{\Phi}, \mathbf{\Psi})$ satisfies the sparsity and incoherence requirements, as described in the previous section, and we focus our attention on the reconstruction module of the sensing system. A concept that lies in the core of CS theory is to use the *a posteriori computing power* to reduce the potential a priori sampling complexity.

Starting with the case of acquiring *noiseless* measurements $\vec{g}$, then, if the RIP holds, the solution of the following optimization problem gives an accurate recovery of $\vec{x}$ via the accurate reconstruction of its corresponding weight vector $\vec{w}$:

$$\vec{w}_{opt} = \arg \min_{\vec{w}' \in \mathbb{R}^N} \|\vec{w}'\|_0 \ , \quad \text{subject to} \quad \vec{g} = \mathbf{\Phi}\vec{w}' \ , \tag{1.11}$$

where $\|\vec{v}\|_0$ denotes the ℓ_0-norm of a vector $\vec{v} \in \mathbb{R}^N$, which is equal to the number of non-zero entries in $\vec{v}$. This optimization problem can recover an L-sparse signal exactly with high probability using only $M = L + 1$ i.i.d. Gaussian measurements [14]. Unfortunately, solving (1.11) is both numerically unstable and NP-complete, requiring an exhaustive enumeration of all $\binom{N}{L}$ possible locations of the nonzero entries in $\vec{w}$.

For this purpose, we relax the above optimization problem by replacing the ℓ_0-norm with the ℓ_1-norm and thus, the solution of the following linear program[3] gives an accurate recovery of $\vec{w}$:

$$\vec{w}_{opt} = \arg \min_{\vec{w}' \in \mathbb{R}^N} \|\vec{w}'\|_1 \ , \quad \text{subject to} \quad \vec{g} = \mathbf{\Phi}\vec{w}' \ , \tag{1.12}$$

where the ℓ_1-norm of a vector $\vec{v} \in \mathbb{R}^N$ is given by $\|\vec{v}\|_1 \triangleq \sum_{i=1}^{N} |v_i|$. That is, among all solutions consistent with the measurements, $\vec{g} = \mathbf{\Phi}\vec{w}'$, we pick that with the minimum ℓ_1-norm. The optimization problem (1.12) is referred to as the ℓ_1-*minimization problem*. This problem may be considered as a *synthesis-based* approach, since we search for the sparsest $\vec{w}_{opt}$ which synthesizes signal $\vec{x}$ (via $\mathbf{\Psi}^{-1}$ ($= \mathbf{\Psi}^T$ in the orthonormal case)). However, it is important to note that we can also follow an *analysis-based* procedure by searching directly for the signal $\vec{x}$ that has a sparse representation in basis $\mathbf{\Psi}$ solving the following problem:

$$\vec{x}_{opt} = \arg \min_{\vec{x}' \in \mathbb{R}^N} \|\mathbf{\Psi}^T\vec{x}'\|_1 \ , \quad \text{subject to} \quad \vec{g} = \mathbf{\Phi}\mathbf{\Psi}^T\vec{x}' \ . \tag{1.13}$$

The use of the ℓ_1-norm as a sparsity-promoting function goes back in the eighties when the ℓ_1-minimization approach was suggested to recover sparse spike trains [15, 16], while recent works [17, 18, 19, 20, 21] study how the minimization of the ℓ_1-norm can recover sparse signals in some special setups. It is also interesting to mention that the problem of recovering L-sparse signals from a small number of frequency samples placed on special equispaced grids, has been carried out from a more computer-science point of view, by employing methods such as isolation and group testing [22].

The following theorem provides strong conclusions with respect to the reconstruction of an L-sparse vector via the solution of (1.12):

[3]The convex optimization problem of minimizing the ℓ_1-norm subject to linear equality constraints can be recast as a linear program.

Theorem 1.2 ([13]) *Assume that the measurement matrix* $\mathbf{\Phi}$ *satisfies the RIP* (1.5) *with* $\epsilon_{2L} < \sqrt{2}-1$. *Then, the solution* $\vec{w}_{opt}$ *to* (1.12) *satisfies*

$$\begin{aligned} \|\vec{w}_{opt} - \vec{w}\|_2 &\leq \frac{c_0}{\sqrt{L}} \cdot \|\vec{w} - \vec{w}_L\|_1 \\ \|\vec{w}_{opt} - \vec{w}\|_1 &\leq c_0 \cdot \|\vec{w} - \vec{w}_L\|_1 \end{aligned} \tag{1.14}$$

for some constant $c_0 > 0$.

Notice that this theorem deals with a broader class of signals, namely, the class of L-compressible signals. If $\vec{w}$ is not strictly L-sparse (that is, it has exactly L non-zero components) but L-compressible (that is, it has only L "significant" components) then, inequality (1.14) asserts that the quality of the reconstructed vector is as good as if we had a prior knowledge of the location of the L largest components of $\vec{w}$ and measured those directly. This is a critical remark, since natural signals are mostly L-compressible rather than L-sparse. Nevertheless, we note again that CS is able to both identify the locations and estimate the amplitudes of the non-zero components without any specific prior knowledge about the signal except its assumed sparsity. For this reason CS is often referred to as a universal approach, since it can effectively recover any sufficiently sparse signal from a set of nonadaptive samples. Optimization based on the ℓ_1-norm can exactly recover L-sparse signals and approximate closely L-compressible signals with high probability using, for instance, only M i.i.d. Gaussian entries, where M satisfies inequality (1.7) [3, 4].

However, it is obvious that in a real-world scenario the sensing system acquires *noisy* measurements, since it will be subjected to measurement inaccuracies:

$$\vec{g} = \mathbf{\Phi}\vec{w} + \vec{n} \ , \tag{1.15}$$

where $\vec{n} \in \mathbb{R}^M$ is a stochastic or deterministic error term (measurement noise) with bounded energy $\|\vec{n}\|_2 \leq \varepsilon$. For the reconstruction of the L-sparse vector $\vec{w}$ the ℓ_1-minimization approach is employed again, but with relaxed inequality constraints:

$$\vec{w}_{opt} = \arg \min_{\vec{w}' \in \mathbb{R}^N} \|\vec{w}'\|_1 \ , \quad \text{subject to} \quad \|\vec{g} - \mathbf{\Phi}\vec{w}'\|_2 \leq \varepsilon \ , \tag{1.16}$$

which means that we only ask the reconstruction be consistent with the data in the sense that the approximation error $\vec{g} - \mathbf{\Phi}\vec{w}'$ be within the noise level. Problem (1.16), which is a convex problem (a second-order cone program (SOCP)) and thus can be solved efficiently, is also known as the LASSO [23].

A variety of reconstruction methods have been proposed to recover an approximation of $\vec{w}$ when the measurements are corrupted by zero-mean Gaussian noise with bounded variance. The solution of the Dantzig selector program [24],

$$\vec{w}_{opt} = \arg \min_{\vec{w}' \in \mathbb{R}^N} \|\vec{w}'\|_1 \ , \quad \text{subject to} \quad \|\mathbf{\Phi}^T(\vec{g} - \mathbf{\Phi}\vec{w}')\|_\infty \leq \kappa_1 \ , \tag{1.17}$$

where $\|\vec{v}\|_\infty = \max_{i=1,\ldots,N} |v_i|$, or the solution to a combinatorial optimization program by performing the following penalized least squares minimization [25],

$$\vec{w}_{opt} = \arg \min_{\vec{w}' \in \mathbb{R}^N} \left\{ \|\vec{g} - \mathbf{\Phi}\vec{w}'\|_2^2 + \kappa_2 \|\vec{w}'\|_0 \right\} \ , \tag{1.18}$$

for appropriately chosen regularization constants κ_1, κ_2 depending on the noise variance, both provide provable reconstruction results.

Similarly to Theorem 1.2, the following statement provides an upper bound for the recon-

struction error in the case of noisy CS measurements:

Theorem 1.3 ([13]) *Assume that the measurement matrix* $\mathbf{\Phi}$ *satisfies the RIP* (1.5) *with* $\epsilon_{2L} < \sqrt{2}-1$. *Then, the solution* $\vec{w}_{opt}$ *to* (1.16) *satisfies*

$$\|\vec{w}_{opt} - \vec{w}\|_2 \leq \frac{c_0}{\sqrt{L}} \cdot \|\vec{w} - \vec{w}_L\|_1 + c_1 \cdot \varepsilon \tag{1.19}$$

for some constants $c_0, c_1 > 0$.

The above reconstruction error is bounded by the sum of two terms. The first is the error which would occur as if we had noiseless data, while the second is just proportional to the noise level.

In this thesis we focus on the development of reconstruction algorithms and thus in the subsequent analysis we will employ the families of matrices that were mentioned before and satisfy the RIP. This is achievable due to the decoupled character of the sensing and reconstruction modules. In particular, for a fixed sensing/encoding part the overall performance of a CS-based acquisition system improves as better reconstruction algorithms are designed and implemented. Before proceeding, we review the main classes of reconstruction algorithms proposed so far in the general case of acquiring noisy CS measurements according to Eq. (1.15).

1.2 Reconstruction techniques for CS

In this section, we review the main categories of CS reconstruction algorithms proposed so far in the literature. For this purpose, we can distinguish two prevalent categories, namely the *deterministic (or norm-based)* reconstruction algorithms, which solve a (convex) norm-based constraint optimization problem, and the *probabilistic (or Bayesian)* reconstruction techniques, which formulate the problem of recovering a sparse signal in a completely probabilistic framework.

1.2.1 Deterministic (norm-based) reconstruction techniques

Some of the CS reconstruction algorithms belonging in this category have been already mentioned, such as the recent works described in [17, 18, 19, 20, 21], the Dantzig selector [24] and the penalized least squares minimization [25], as well as the works presented in [26, 27, 28]. Other algorithms solving an ℓ_1-minimization problem include interior-point [29, 30] and fixed-point continuation methods [31], which are slow in general but very accurate, homotopy methods [32, 33], which are fast and accurate especially for small-scale problems, and gradient projection algorithms [34, 35, 36], which are fast and can efficiently handle large-scale problems.

Greedy algorithms can be also very efficient and computationally tractable when the signal of interest is highly sparse. In this family belong algorithms such as the Basis Pursuit (BP) [37] and its variants, namely, the Orthogonal Matching Pursuit (OMP) [38], the stagewise OMP (StOMP) [39], the regularized OMP (ROMP) [40] and the thresholded BP [41]. Additionally, there are also sparse signals whose non-zero coefficients occur in clusters, the so-called block-sparse signals. For the reconstruction of these signals, a block version of the OMP algorithm (BOMP) [42] and a mixed ℓ_1/ℓ_2 optimization approach [43, 44] have been proposed, which also provide the sufficient conditions on block-coherence to guarantee the recovery of block L-sparse signals through BOMP.

Iterative thresholding techniques [45, 46, 47, 48] constitute another family of deterministic reconstruction approaches, which are very fast and recover sparse signals very accurately, while they also perform well for recovering compressible signals.

1.2.2 Probabilistic (Bayesian) reconstruction techniques

The deterministic reconstruction approaches reviewed in the previous section solve a constrained-based optimization problem resulting in a *point estimate* of the sparse signal, the weight vector $\vec{w}$ in our case. Bayesian approaches have been widely reported in the field of sparse Bayesian learning [49, 50, 51, 52].

The development of CS reconstruction algorithms in a probabilistic/Bayesian framework is still in a relatively early stage, and only in a few recent studies [53, 54, 55, 56] the inversion of CS measurements was considered from such a perspective. In particular, given a prior belief that the weight vector $\vec{w}$ should be sparse in basis $\mathbf{\Psi}$ and the set of CS measurements $\vec{g}$, the objective is to formulate a *posterior probability distribution* for $\vec{w}$. This improved the accuracy over the point estimates and provided confidence intervals (error bars) in the approximation of the original signal $\vec{x}$. Besides, this was also used to guide the optimal design of additional CS measurements implemented with the goal of reducing the uncertainty in reconstructing $\vec{x}$ [57].

In this probabilistic framework, the assumption that $\vec{w}$ is sparse is formalized by modeling the distribution of $\vec{w}$ using a sparseness-enforcing prior distribution. A common choice of this prior is the Laplace density [58]. However, the use of a Laplace prior density raised the problem that the Bayesian inference could not be performed in closed form, since the Laplace prior is not conjugate[4] to the Gaussian assumption made for the likelihood model. For this purpose, a hierarchical prior model was invoked using a set of hyperparameters, which had similar properties as the Laplace prior but allowed convenient conjugate-exponential analysis. Then, the overall prior on $\vec{w}$ was evaluated analytically, resulting in the the Student-t distribution [50], which can be considered as a sparseness prior, since it is peaked about zero. An efficient method for estimating the corresponding model parameters is described in [59].

A fast tree-structured matching pursuit algorithm developed in the Bayesian framework is presented in [60], while [61] presents an accelerated CS encoding/decoding algorithm that employees belief propagation by emphasizing a two-state mixture Gaussian model as a prior for sparse signals.

The majority of the contributions which appear in the literature address the recovery of a single sparse signal. However, there are cases in several concrete applications where an ensemble of (multi-channel) signals may not only possess sparse expansions for each signal (channel) individually, but additionally the distinct signals can also exhibit common sparsity structures. Motivated from previous studies carried out in a machine-learning perspective and the related research in the field of simultaneous sparse approximation (SSA) [62, 63, 64, 65], the BCS framework was extended recently to the so-called multi-task CS [66, 67] for the simultaneous reconstruction of an ensemble of sparse signals, presenting a common sparsity structure, using multiple CS measurements.

From a CS point of view, the ideas presented under the multi-task perspective, are extended in the framework of *distributed compressed sensing*, which is particularly suitable for networked data, as it is reviewed in section.

1.3 Distributed CS

Although the theory and implementation of compression have been well developed for single signals, however, many applications involve multiple signals, for which there has been less progress. For instance, consider a sensor network in which a number of distributed nodes acquire data and transmit them in a central collection point, the so-called *fusion center* (FC) [68]. In such networks, communication energy and bandwidth are often scarce resources, making the reduction

[4] In probability theory, a family of prior probability distributions $p(s)$ is said to be conjugate to a family of likelihood functions $p(x|s)$ if the resulting posterior distribution $p(s|x)$ is in the same family as $p(s)$.

of communications critical. Fortunately, since the sensors usually observe related phenomena, the ensemble of signals they acquire is expected to possess some joint structure, or inter-signal correlation in addition to the intra-signal correlation in each individual sensor's measurements. For example, the sensor network data often contain spatial and temporal correlations. Such structures can be exploited by *distributed source coding* algorithms, where each signal is encoded separately and all signals are recovered jointly allowing a substantial savings on the communication cost [69, 70, 71]. The CS recovery algorithms mentioned so far are designed mainly to exploit intra-signal correlation structures at a single sensor.

Unfortunately, practical schemes for distributed compression of sources with both types of correlation have remained a challenging problem for quite some time. The theory of *Distributed Compressed Sensing* (DCS) [14] enables new distributed encoding/sensing algorithms that exploit both intra- and inter-signal correlation structures. In a typical DCS scenario, a number of sensors acquire signals that are each individually sparse in some transform basis and also correlated from sensor to sensor [72]. Each sensor *independently encodes* its signal by projecting it onto another incoherent measurement basis (cf. Section 1.1.1), with each sensor operating entirely without collaboration, and then transmits just a few of the resulting coefficients to the FC. Under the right conditions, a decoder at the FC can *reconstruct jointly* all of the signals precisely. The DCS theory relies on the concept of *joint sparsity* of a signal ensemble [14, 73, 74].

Like the CS framework, the DCS approach is also *"democratic"* in the sense that the randomized measurements coming from each sensor have equal priority and they carry the same amount of information. This results in sensing schemes which are robust to measurements loss and quantization, as well as they allow a progressively better reconstruction of the data as more measurements become available [75]. The second critical property of a DCS scheme (similarly to the CS one) is its *asymmetrical* nature, since the encoding process is very simple; the sensors merely compute incoherent projections with their signals, while it places most of the computational complexity in the joint decoder, which often has more substantial resources than any individual sensor node. The goal when designing a DCS method is to minimize the overall sensor measurement rates in order to reduce the total communication cost.

The performance of a DCS reconstruction scheme when applied on networked data, that is, on a set of observations gathered by the sensors of a network, is highly dependent on the selection of a suitable sparsifying basis $\mathbf{\Psi}$. But, while transform-based compression is well developed in traditional signal and image processing tasks, the design of appropriate sparsifying bases for networked data is still in an early stage. A few recent approaches associate a graph with a given network, where the vertices of the graph represent the nodes of the network and edges between vertices represent anticipated relationships among the data at adjacent nodes, such as communication links or correlations and dependencies between data.

If the nodes are placed in a uniform fashion, then the underlying graph can be represented as a regular lattice. In this setting, the sensor locations can be viewed as sampling locations and methods like the discrete Fourier transform (DFT) or the discrete wavelet transform (DWT) may be used to sparsify the sensor data. In more general random settings, wavelet techniques can be extended to also handle the irregular distribution of sampling locations [76]. In particular, graph wavelets were developed by adapting the design principles of the DWT to more abstract graph topologies [77], while the so-called diffusion wavelets provide an alternative approach by constructing an orthonormal basis (in contrast to graph wavelets that produce an overcomplete dictionary) tailored to functions supported on a specific graph [78].

Apart from applying CS in several network-based monitoring applications, this sampling framework could be also effective in other signal and image processing applications, as it is briefly reviewed in the next section.

1.4 Applications of CS

The class of compressible signals appears in many diverse fields and thus making CS a suitable tool for capturing efficiently the salient information using a set of incoherent measurements with cardinality proportional to the signal's *information level* (sparsity), which is significantly smaller than its size. In the following, we review briefly some applications which were of the first to employ the advantages of CS theory.

- *Signal detection-classification*: The CS literature has mainly focused on problems in signal reconstruction, approximation, and estimation in the presence of noise. Recent studies [79, 80, 81, 82] reveal that the CS framework is *information scalable* to a wider range of statistical inference tasks, such as detection and classification of a signal (or target), which do not require a reconstruction of the signal, but only estimates of the relevant sufficient statistics for the problem at hand. As a consequence, these works showed that significantly fewer measurements are required for signal detection and classification than for signal reconstruction, while the computational complexity is also much reduced. The problem of target localization belongs in this broader class of applications. The standard, as well as the distributed CS framework have been employed to exploit the sparsity of the received signal emitted by a target in a suitable domain [83, 84].

- *Compressive imaging*: The theory of CS has been extended in the case of processing digital image or video, resulting in the theory of compressive imaging (CI). The combination of mathematical and computational methods could have a great impact in this area, where conventional hardware design has significant limitations. For instance, conventional imaging devices that use CCD or CMOS technology are limited essentially to the visible spectrum. However, a CI camera that collects incoherent measurements using a digital micromirror array (and requires just one photosensitive element instead of millions) could significantly expand these capabilities [85, 86, 87].

- *Medical/Astronomical imaging*: The fields of medical and astronomical image processing are two fields that exploit the advantages of CS, each one for its own reasons. For instance, magnetic resonance imaging (MRI) obeys two key requirements for successful application of CS: i) medical imagery is compressible by sparse coding in an appropriate transform domain (e.g., by applying the 2-D wavelet transform) and ii) MRI scanners naturally acquire encoded samples, rather than direct pixel samples (e.g., in spatial-frequency encoding). Of particular interest is the way in which different applications face different constraints, imposed by MRI scanning hardware or by patient considerations, and how the inherent freedom of CS to choose sampling trajectories and sparsifying transforms plays a key role in matching the constraints [88, 89].

  On the other hand, due to particular data acquisition procedures, such as raster scans, the astronomical data are often redundant. This fact, in conjunction with several technical challenges that stem primarily from the constraints imposed by the sensing and processing devices, such as limited power, communication bandwidth, and small storage capacity, motivate the need for a new paradigm for data processing. In this case, CS enables a potentially significant reduction in the sampling and computation costs at an on-board sensing device with limited capabilities [90]. Besides, the asymmetrical nature of CS-based approaches is a crucial point for the design of on-board sensing devices. In particular, the property of asymmetry refers to the fact that the compression/encoding part is of very low computational cost (simple linear random projections), while the main computational burden is on the decompression/decoding part (for instance, a base-station on the earth), where increased processing capabilities and computational resources are available. In

addition, CS is able to account for the redundancy of the data during the compression stage due to particular acquisition processes.

- *Communications*: The CS principles (sparsity, randomness and convex optimization) have been applied to design fast and efficient error correcting codes over the reals to protect from errors during transmission, as well as for channel estimation and equalization purposes [93, 94]. Analog-to-Information (A/I) conversion is also an emerging field in which a discrete, low-rate sequence of incoherent measurements can be acquired from a high-bandwidth analog signal with the goal of alleviating the pressure on conventional ADC technology, which is currently limited to sample rates on the order of 1 GHz [95, 96]. Working in the discrete CS framework, the experimental results suggest that analog signals obeying a sparse or compressible model (in some analog dictionary $\mathbf{\Psi}$) can be captured efficiently using these devices at a rate proportional to their information level instead of their Nyquist rate.

1.5 Contributions of the thesis

The field of compressed sensing signal processing has gained the increasing interest of the research community resulting in the development of a vast number of reconstruction methods, many of which were referred in previous sections, in the last few years only. However, there will be always space for improvements.

The main contributions of this thesis concern the development of novel Bayesian techniques at the decoding/reconstruction part of a CS-based system. In particular, we go against the common tenet of using hierarchical models for modelling the sparsity of a given signal, or the trend of making a Gaussian assumption for the statistics of the sparse vector and/or the noise. For this purpose, all of the proposed CS reconstruction algorithms model the (possibly highly) sparse behavior of a given signal by employing directly an appropriate sparsity-enforcing univariate or multivariate distribution on the components of the sparse vector. Besides, in all of the proposed methods we assume highly impulsive environments, which cannot be modelled accurately under a Gaussian assumption. For this purpose, we employ heavy-tailed models which are suitable in capturing the true statistical characteristics of such signals.

The main contributions of this thesis can be summarized as follows:

- The class of Bayesian CS reconstruction methods, which is still restricted in comparison with the class of norm-based approaches, is enriched with the development of a method that employs a Gaussian Scale Mixture (GSM) , which also comprises some of the recently introduced Bayesian CS algorithms as special cases.

- The problem of reconstructing a signal based on a set of multiple observation vectors was mainly treated by norm-based optimization approaches. In this thesis, we present an extension of the GSM-based method that is able to handle multiple observations, generated by projecting a single original signal on a set of over-complete dictionaries, in a distributed manner and in a probabilistic framework and thus exploiting the advantages offered by a Bayesian inference procedure.

- The lack of closed-form expressions, which is usually present in most of the Bayesian inference tasks when using an assumption other than the Gaussian, is the main reason of not using other statistical models, which may provide closer approximations to the true underlying statistical characteristics of the signal of interest. As a result, the vast majority of the existing CS algorithms employs the Gaussian distribution for modeling the signal and/or the noise components, which may be inaccurate in many real-world applications,

such as in underwater acoustics and telecommunications. We overcome this limitation by proposing a CS method for reconstructing highly impulsive signals in the presence of heavy-tailed noise, by employing the Cauchy as their prior distribution. A greedy Bayesian approach combined with a tree-structure also results in a fast implementation with closed-form expressions.

- Although the family of *alpha-Stable* distributions is extremely powerful for representing highly impulsive and thus sparse phenomena in the context of CS, however, the lack of closed-form expressions and second-order statistics make it difficult to manipulate. The last of our proposed CS methods can be considered as a generalization of the two previous methods in the sense that it is able to handle a general heavy-tailed model, as it is the case of a *Symmetric alpha-Stable* ($S\alpha S$) distribution. For instance, the Cauchy distribution employed by the previous method belongs in this family.

- We overcome the non-adaptive behavior inherent to any CS method by introducing a new $S\alpha S$ measurement matrix, which is adapted to the statistical characteristics of the sparse signal.

- A novel Lagrangian function is introduced for the estimation of a sparse vector by solving a constrained optimization problem using the duality theory and the method of subgradients. The proposed Lagrangian is best adapted to the case of *alpha-Stable* distributions by exploiting *Fractional Lower-Order Moments* instead of second-order statistics, which are not defined for this family of distributions.

Part I

Mathematical Background

Chapter 2

Wavelet Transform - a sparsity enforcing domain

Before a diamond shows its brilliancy and prismatic colors it has to stand a good deal of cutting and smoothing.

Anonymous

2.1 Introduction

In the following two chapters we review briefly the main mathematical tools which will be employed in the subsequent analysis of the proposed CS reconstruction techniques. As mentioned in Chapter 1, one of the fundamental properties of a CS-based system is its decoupling behavior with respect to the compression and the reconstruction modules. Figure 2.1 shows the compression module[1] of such a system. The core tenet of CS theory requires that the signal of interest is sparse by itself (in the space-domain) or in a suitable transform basis. Thus, we are interested in transformations that result in as highly sparse coefficient vectors as possible (in the discrete case).

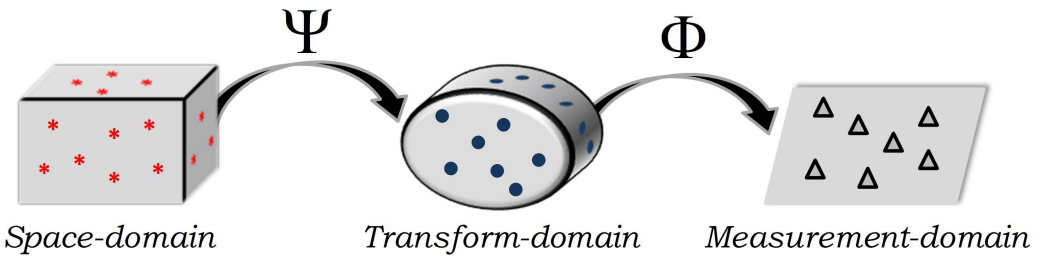


Figure 2.1: The compression/encoding module of a CS-based system.

Several works [6, 7] have shown that the use of a wavelet basis in place of $\mathbf{\Psi}$ yields a highly compact (sparse) representation of many natural signals. Since the experimental evaluation of our proposed CS reconstruction methods will be primarily exploit the efficiency of the wavelet transform we cite here its main concepts and properties. The above references can be used for a more thorough study of the subject.

[1] In the subsequent text the terms compression/encoding and decompression/decoding/reconstruction will be used interchangeably, respectively.

2.2 Discrete Fourier Transform (DFT)

Fourier analysis is one of the oldest subjects in mathematics and it is of great importance to both mathematicians and engineers, since it has many applications in different fields. Fourier analysis is well known for the (integral) Fourier transforms and Fourier series of a function $f(t)$ defined on the real line $\mathbb{R}$. When $f(t)$ represents an analog signal, then it is defined in the continuous-time domain, while its Fourier transform describes the spectral behavior of $f(t)$ in terms of frequency. On the other hand, a Fourier series expansion is a transformation of bi-infinite sequences to periodic functions.

For each p, $1 \leq p < \infty$, let $L^p(\mathbb{R})$ denote the class of measurable functions on $\mathbb{R}$ such that the integral

$$\int_{-\infty}^{\infty} |f(t)|^p \, \mathrm{d}t$$

is finite ($< \infty$). Since the continuous Fourier transform is used to describe the spectral content of an analog signal $f(t)$ with finite energy (that is, $f(t) \in L^2(\mathbb{R})$), we are also interested in describing the spectral content in the case of a digital (discrete-time) signal. A digital signal is represented by a sequence $\{f[n] = f(nT)\}_{n\in\mathbb{Z}} \in \ell^p(\mathbb{Z})$ sampling the values of the corresponding analog signal with period T, where $\ell^p(\mathbb{Z})$ denotes the spaces of bi-infinite sequences $\{a_n\}_{n\in\mathbb{Z}}$ with finite l_p-norm, $\|\{a_n\}\|_{l_p} < \infty$, where

$$\|\{a_n\}\|_{l_p} = \begin{cases} \left(\sum_{n\in\mathbb{Z}} |a_n|^p\right)^{\frac{1}{p}}, & 1 \leq p < \infty, \\ \sup_n |a_n| \quad , & p = \infty. \end{cases}$$

The spectral content of a digital signal is extracted by employing the Discrete Fourier Transform (DFT) $\mathcal{F}^*$ as follows:

$$(\mathcal{F}^*\{f[n]\})(u) = \sum_{n\in\mathbb{Z}} f[n] e^{inu} \, . \tag{2.1}$$

In other words, the DFT of $\{f[n]\}_{n\in\mathbb{Z}}$ is the Fourier series with Fourier coefficients given by $\{f[n]\}_{n\in\mathbb{Z}}$. The values of an N-sample time-domain sequence $f[n]$, $n = 0, \ldots, N-1$, and the corresponding N-point transform sequence $F[k]$, $k = 0, \ldots, N-1$, discrete spectrum are given by:

(***Discrete Fourier Transform***)

$$F[k] = \sum_{n=0}^{N-1} f[n] e^{-i2\pi kn/N} \, , \quad k = 0, \ldots, N-1$$

(***Inverse Discrete Fourier Transform***)

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] e^{i2\pi kn/N} \, , \quad n = 0, \ldots, N-1 \, .$$

2.2.1 Short-Time Fourier Transform (STFT)

DFT is useful only when the signal is stationary. It can be employed in the non-stationary case (when the signal is composed of time-varying spectral characteristics), but only if we are interested in what frequency components exist in the signal and not about the corresponding time instants in which they occur. Thus, if we are interested in acquiring information both in the time and frequency domain for a non-stationary signal then, the DFT is not adequate.

The solution to this problem is given by a *two-dimensional* time-frequency representation $S_f(\tau, \omega)$ of $f(t)$ by observing the signal through a "window" $g(t)$ of limited extend centered at time location τ (with frequency ω), in which it is reasonable to assume that the portion of the signal which is seen through it is stationary. More specifically, in the discrete-time domain the short-time Fourier transform (STFT) of $f[n]$ is the DFT of the windowed signal $f[n]g[n-m]$,

$$S_f(m,k) = \sum_{n=0}^{N-1} f[n]g[n-m]e^{-i2\pi kn/N} \ , \ m = 0, \ldots, N-1, \ k = 0, \ldots, N-1 \ , \tag{2.2}$$

which maps f in a time-frequency plane (m, k). Similarly to the inverse DFT, the *inverse STFT* is defined as:

$$f[n] = \frac{1}{E_g} \sum_{k=0}^{N-1} \sum_{m=0}^{N-1} S_f(m,k)g[n-m]e^{i2\pi kn/N} \ , \ n = 0, \ldots, N-1 \ , \tag{2.3}$$

where E_g is a normalizing factor depending on the "energy" of the window g. One of the inherent characteristics of STFT is that it has the same resolution across the time-frequency plane as shown in Figure 2.2(a).

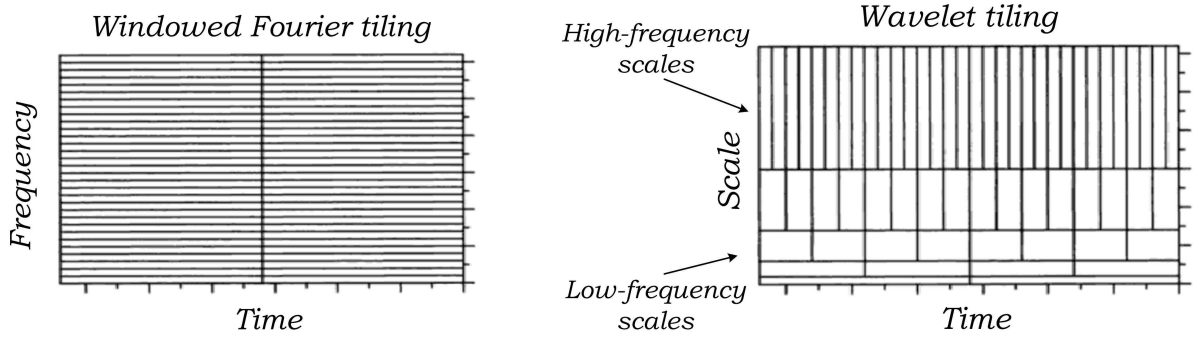


Figure 2.2: Comparison of time-frequency tiling for STFT and wavelet spectrum.

2.3 Wavelet Transform

As we saw above, the main drawback of the STFT is the fixed time-frequency resolution over the entire time-frequency plane, since STFT uses the same window at all frequencies. On the other hand, since frequency is directly proportional to the number of cycles per unit time, it takes a narrow window in the time-domain to locate high-frequency components more precisely and a wide time-window to analyze low-frequency behavior. Hence, the STFT is not suitable for analyzing signals with both very high and very low frequencies. The resolution limitation of the STFT is overcome by introducing the *wavelet transform*, which provides a flexible window which narrows when observing high-frequency components and widens when observing low-frequency components.

2.3.1 The 1-Dimensional Wavelet Transform

Continuous Wavelet Transform (CWT)

Starting with the continuous-time case, the definition of the 1-D CWT is stated as follows:

Definition 2.1 (1-D Continuous Wavelet Transform (CWT))
If $\psi(t) \in L^2(\mathbb{R})$ satisfies the "admissibility" condition

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|}\,\mathrm{d}\omega < \infty\ , \tag{2.4}$$

then $\psi(t)$ is called a basic wavelet. The CWT of a function $f(t)$, relative to a basic wavelet $\psi(t)$, is defined by

$$(W_\psi f)(b,\,\alpha) = \frac{1}{\sqrt{|\alpha|}} \int_{-\infty}^{\infty} f(t)\overline{\psi\Big(\frac{t-b}{\alpha}\Big)}\,\mathrm{d}t, \quad f \in L^2(\mathbb{R})\ , \tag{2.5}$$

where $\alpha,\, b \in \mathbb{R}$ $(\alpha \neq 0)$ and $\Psi(\omega)$ is the Fourier transform of $\psi(t)$,

where $\overline{\psi(\cdot)}$ denotes the conjugate of the function $\psi(\cdot)$, b represents the time-location parameter and α the scale parameter which is inversely proportional to the frequency, $\alpha = \frac{\omega_0}{\omega}$, with ω_0 denoting the central frequency of $\Psi(\omega)$ [6, 97]. By setting

$$\psi_{b,\,\alpha}(t) = \frac{1}{\sqrt{|\alpha|}}\,\psi\Big(\frac{t-b}{\alpha}\Big)\ , \tag{2.6}$$

the CWT becomes

$$(W_\psi f)(b,\alpha) = \int_{-\infty}^{\infty} f(t)\overline{\psi_{b,\,\alpha}}(t)\,dt = \langle f, \psi_{b,\,\alpha}\rangle\ , \tag{2.7}$$

which shows that the wavelet transform gives a measure of similarity, in the sense of similar frequency content, between the basis functions (wavelets) and the signal itself. The basic wavelet, which corresponds to the scale[2] $\alpha = 1$, is rescaled by dilation or compression in order to extract the signal's spectrum at frequencies other than the frequency of the basic wavelet. This is one of the main differences with STFT, which uses an analysis window that contains a number of modulation frequencies. The second main difference is that the STFT uses an analysis window of fixed length at all frequencies, while the CWT uses an analysis window (the basic wavelet) with a length which depends on the frequency, as shown in Figure 2.2(b) on the preceding page.

Finally, we can reconstruct any finite-energy signal $f(t) \in L^2(\mathbb{R})$ from its CWT values using the following inverse formula,

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{(W_\psi f)(b,\alpha)\}\,\psi_{b,\,\alpha}(t)\,\frac{\mathrm{d}\alpha\,\mathrm{d}b}{\alpha^2}\ . \tag{2.8}$$

Note that the finiteness of the constant C_ψ (Eq. (2.4)) restricts the class of $L^2(\mathbb{R})$ functions $\psi(t)$ that can be used as basic wavelets. For a more detailed study on the properties of the CWT the interested reader is referred to [6, 97].

In practice we are interested in reconstructing a signal by sampling it on a discrete set of the time-scale plane (that is, $b, \alpha \notin \mathbb{R}$ anymore). For $\psi_{b,\,\alpha}(t)$ to cover the whole time-axis at a fixed discretized scale $\alpha = \alpha_0^m$, we have to choose the translation parameter $b = nb_0\alpha_0^m$,

$$\alpha = \alpha_0^m,\ b = nb_0\alpha_0^m, \quad m, n \in \mathbb{Z},\ \alpha_0 > 1,\, b_0 > 0\ ,$$

obtaining a discretized family of wavelets

$$\psi_{n,\,m}(t) = \alpha_0^{-m/2}\,\psi(\alpha_0^{-m}t - nb_0) \tag{2.9}$$

[2] In the subsequent analysis the terms scale and level will be used interchangeably.

(the normalization factor $\alpha_0^{-m/2}$ makes $\psi_{n,m}(t)$ of unit norm). The case corresponding to $\alpha_0 = 2$ and $b_0 = 1$ is called *dyadic* and then we can find orthonormal bases which are used in the reconstruction of a function from its CWT coefficients.

When this discretized set of wavelet functions constitutes an orthonormal basis of $L^2(\mathbb{R})$ we can reconstruct a signal $f(t) \in L^2(\mathbb{R})$ using its CWT coefficients $\langle f(t), \psi_{n,m}(t)\rangle$ in a wavelet series expansion,

$$f(t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \langle f(t), \psi_{n,m}(t)\rangle \, \psi_{n,m}(t) \; . \tag{2.10}$$

Instead of using orthonormal wavelet families, $f(t)$ can be reconstructed using bi-orthogonal families of wavelets. In this case the wavelet used for the analysis is different from the one used for the reconstruction [6].

Multi-resolution analysis

A fundamental concept in wavelet analysis is its *multi-resolution* behavior. It is based on the idea of signal decompositions which rely on successive approximations. The given signal will be decomposed in a coarse approximation plus added details. By applying the successive approximations recursively the space of signals $L^2(\mathbb{R})$ can be spanned by spaces of successive details at all resolutions. This happens due to the fact that, as the detail resolution goes to infinity, the approximation error goes to zero. We have the following definition [97]:

Definition 2.2 (Multi-resolution analysis) *A multi-resolution analysis is a sequence $(V_j)_{j\in\mathbb{Z}}$ of closed subspaces of $L^2(\mathbb{R})$ such that*

1. $\ldots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \ldots$
2. $\overline{\bigcup}_{j\in\mathbb{Z}} V_j = L^2(\mathbb{R})$
3. $\bigcap_{j\in\mathbb{Z}} V_j = \emptyset$
4. $f(t) \in V_j \Leftrightarrow f(2^j t) \in V_0$
5. $f(t) \in V_0 \Leftrightarrow f(t-n) \in V_0 \; , \forall n \in \mathbb{Z}$
6. *there exists a function $\varphi(t) \in V_0$, called a scaling function, such that the set $\{\varphi(t-n)\}_{n\in\mathbb{Z}}$ is an orthonormal basis in V_0.*

Note that by combining *4.-6.* we conclude that the set $\{\varphi_{n,j}(t) = 2^{-j/2}\varphi(2^{-j}t - n)\,,\, n \in \mathbb{Z}\}$ constitutes a basis for V_j.

Now consider W_j to be the orthogonal complement of V_j in V_{j-1}, that is, $V_{j-1} = V_j \oplus W_j$, which leads to the decomposition

$$L^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j \; . \tag{2.11}$$

Besides, using the property *2.* of the V_j spaces an analogous relation holds for the W_j spaces,

$$f(t) \in W_j \Leftrightarrow f(2^j t) \in W_0 \; . \tag{2.12}$$

Hence, the translations of the basic wavelet $\psi(t) \in W_0$, $\{\psi(t-n)\,,\, n \in \mathbb{Z}\}$, form an (orthonormal) basis of W_0, whereas from properties *2.-3.* and Eq. (2.11) the set

$$\psi_{n,j}(t) = 2^{-j/2}\psi(2^{-j}t - n), \quad n,\, j \in \mathbb{Z} \tag{2.13}$$

constitutes an orthonormal basis of $L^2(\mathbb{R})$, while $\{\psi_{n,j}\}_{n\in\mathbb{Z}}$ is an orthonormal basis of W_j. The scaling function $\varphi(t)$ has the characteristics of a lowpass filter, while the wavelet function $\psi(t)$ has the characteristics of a highpass filter.

Concluding, the following interpretation can be given to the multi-resolution analysis. From property *2.*, every signal $f \in L^2(\mathbb{R})$ can be approximated arbitrarily closely by an $f_k \in V_k$, for some $k \in \mathbb{Z}$. Since $V_{j-1} = V_j \oplus W_j$, f_k has the following decomposition:

$$f_k = f_{k+1} + g_{k+1} \ ,$$

where $f_{k+1} \in V_{k+1}$ is a coarsest approximation in the "approximation space" V_{k+1} and $g_{k+1} \in W_{k+1}$ is a detail in the "detail space" W_{k+1}. By repeating this process we obtain

$$f_k = g_{k+1} + g_{k+2} + \cdots + g_{k+P} + f_{k+P}$$

where $f_j \in V_j$ and $g_j \in W_j$. In general we can represent a signal $f(t)$ as a succession of multi-resolution details and a coarsest approximation. Then, the original signal f can be recovered by following the inverse process, that is, by adding the details to the coarsest approximation.

Discrete Wavelet Transform (DWT)

By discretizing the continuous wavelet transform over the set of indices (n, m) we obtain a discrete set of wavelet functions $\psi_{n,m}$, which perform the analysis and synthesis of a signal $f(t)$. The analysis coefficients are given by the following projections,

$$c_{n,m} = \langle f, \psi_{n,m} \rangle$$

and from these the original signal is recovered as follows,

$$f(t) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} c_{n,m}\, \psi_{n,m}(t) \ .$$

The implementation of the *Discrete Wavelet Transform* (DWT) assumes the existence of a multi-resolution analysis produced by a scaling function $\varphi(t)$.

Since the approximation spaces V_j are getting larger as $j \to -\infty$, $f(t)$ can be approximated arbitrarily closely by choosing a small enough $j = -J$ $(J > 0)$ and taking the projection of this function into V_{-J} using the basis functions $\{\varphi_{n,-J}(t)\}_{n\in\mathbb{Z}}$. The approximation coefficients are given by

$$a_0[n] = \langle f(t), \varphi_{n,-J}(t) \rangle \ ,$$

which can be employed to reconstruct an approximation of $f(t)$,

$$f(t) \approx \sum_{n=-\infty}^{\infty} a_0[n] \varphi_{n,-J}(t) \ .$$

Thus, the original signal $f(t)$ can be represented by its approximation in V_{-J} (or equivalently at *level-0*) via the set of coefficients $a_0[n]$. By exploiting the property $V_{j-1} = V_j \oplus W_j$ the above approximation can be decomposed further using the basis functions $\{\varphi_{n,-J+1}(t)\}_{n\in\mathbb{Z}}$ in V_{-J+1} and $\{\psi_{n,-J+1}(t)\}_{n\in\mathbb{Z}}$ in W_{-J+1} resulting in the following expansion,

$$\begin{aligned} f(t) &= \sum_{n=-\infty}^{\infty} a_0[n]\,\varphi_{n,-J}(t) \\ &= \sum_{n=-\infty}^{\infty} a_1[n]\,\varphi_{n,-J+1}(t) + \sum_{n=-\infty}^{\infty} d_1[n]\,\psi_{n,-J+1}(t) \\ &= A_1(t) + D_1(t) \end{aligned}$$

where $A_1(t)$ corresponds to the *approximation* at *level-1* and $D_1(t)$ denotes the *detail* at the same level, while the coefficients $a_1[n]$ and $d_1[n]$ are the approximation and detail coefficients of the first level, respectively. This decomposition procedure can be repeated for the approximation $A_j(t)$ at *level-j* resulting in an iterative expansion,

$$\begin{aligned} f(t) &= A_1(t) + D_1(t) \\ &= A_2(t) + D_2(t) + D_1(t) \\ &= A_3(t) + D_3(t) + D_2(t) + D_1(t) \\ &= \ldots. \end{aligned}$$

This decomposition procedure is equivalent to a lowpass (*Lo-D*) and highpass (*Hi-D*) filtering followed by downsampling, as Figure 2.3[3] shows.

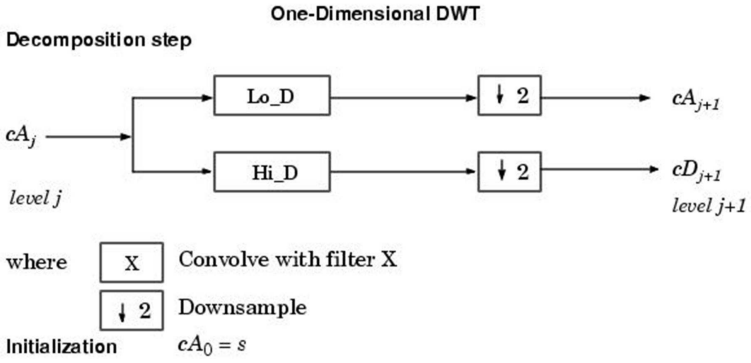


Figure 2.3: The filtering process of the 1-D DWT.

2.3.2 The 2-Dimensional Discrete Wavelet Transform

An extension of the 1-D DWT is available for the class of two-dimensional (2-D) discrete signals, such as the digital images. In two dimensions, a 2-D scaling function $\varphi(x, y)$ and three 2-D directionally sensitive wavelets, $\psi^H(x, y)$, $\psi^V(x, y)$, $\psi^D(x, y)$ are required. These functions can be written in a separable form in terms of a 1-D scaling function $\varphi(t)$ and its corresponding basic wavelet $\psi(t)$, as follows:

$$\varphi(x, y) = \varphi(x)\varphi(y) \tag{2.14}$$
$$\psi^H(x, y) = \psi(x)\varphi(y) \tag{2.15}$$
$$\psi^V(x, y) = \varphi(x)\psi(y) \tag{2.16}$$
$$\psi^D(x, y) = \psi(x)\psi(y) \tag{2.17}$$

[3]Obtained from http://www.mathworks.com/access/helpdesk/help/toolbox/wavelet/dwt.html

The directionally sensitive character of these wavelets expresses their ability to measure functional variations (e.g., intensity or gray-level variations for images) along different directions. So, $\psi^H(x,y)$ measures variations along the horizontal direction, $\psi^V(x,y)$ localizes variations along the vertical direction and $\psi^D(x,y)$ responds to variations along diagonals. Given the above separable functions, the extension of the 1-D DWT to two dimensions is direct. In the 2-D case the basis functions at scale j are defined as follows [98],

$$\varphi_{j,m,n}(x,y) = 2^{j/2}\varphi(2^j x - m, 2^j y - n) \tag{2.18}$$

$$\psi^i_{j,m,n}(x,y) = 2^{j/2}\psi^i(2^j x - m, 2^j y - n)\ , \qquad i \in \{H,V,D\} \tag{2.19}$$

Definition 2.3 (2-D Discrete Wavelet Transform) *The 2-D DWT of a signal $f(x,y)$ of size $M \times N$ at a fixed scale j, is defined as follows:*

$$A_j(m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\varphi_{j,m,n}(x,y) \tag{2.20}$$

$$D^i_j(m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\psi^i_{j,m,n}(x,y)\ ,\ i \in \{H,V,D\} \tag{2.21}$$

where A_j denotes the set of approximation wavelet coefficients and D^i_j the set of detail coefficients across direction $i \in \{H,V,D\}$, at the given scale j (these four sets of coefficients constitute the *subbands* of the 2-D DWT for the given signal $f(x,y)$). Following a similar approach as in the 1-D case by decomposing iteratively the approximation coefficients at each level we obtain the structure shown in Figure 2.4 (in the case of 3 levels).

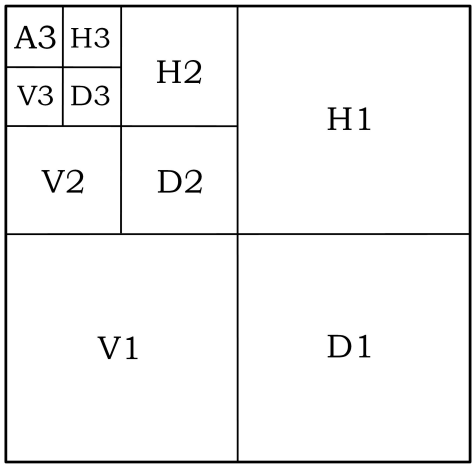


Figure 2.4: A 3-level 2-D DWT.

Figure 2.5 on the next page shows the original image "Facets" along with its 3-level 2-D DWT using the Daubechies' 4 ("db4") wavelet. The normalized histogram of the amplitudes of the detail coefficients at all decomposition levels reveals the sparsity enforcing behavior of the 2-D DWT, since the vast majority of wavelet coefficients is concentrated around zero.

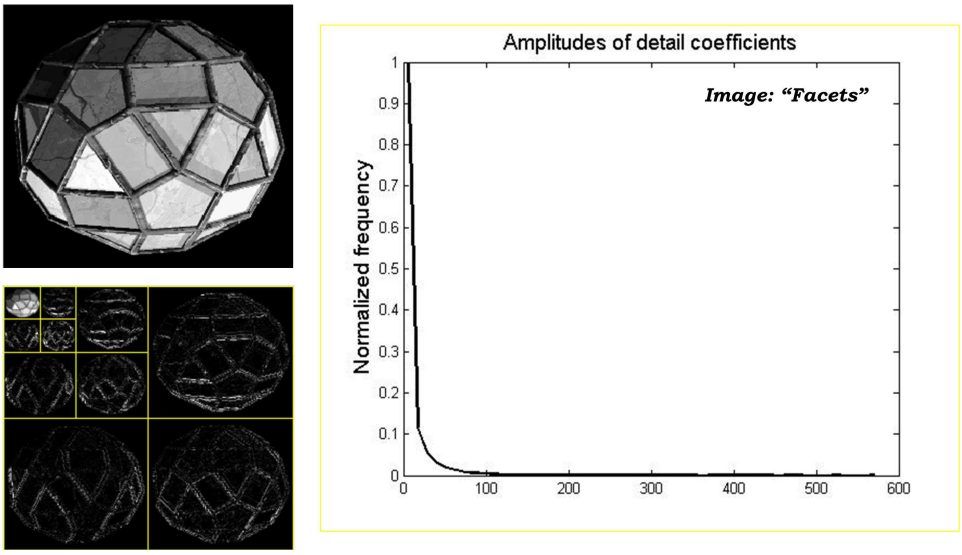


Figure 2.5: Original image "Facets", along with its 3-level 2-D DWT using Daubechies' 4 wavelet.

2.4 Statistical properties of the DWT coefficients

The development of CS reconstruction methods in a Bayesian framework, as it will be described in detail in the subsequent chapters, will be more convenient when the coefficients of a properly chosen wavelet transform satisfy certain statistical properties. Fortunately, the de-correlating capability of the DWT is guaranteed and it can be verified by evaluating the auto- and cross-correlation of the wavelet coefficients among the several subbands. A detailed study on the correlation properties of these coefficients [99] showed that the statistical expectation of the approximation coefficients is proportional to the average of the original signal, whereas the corresponding expectation of the detail coefficients is almost zero, inducing a high sparsity of the original signal in the wavelet domain. It is also proved that the auto-correlations of the approximation and detail coefficients are proportional to the auto-correlation of the original signal. On the other hand, the cross-correlation between approximation and detail coefficients at the same resolution level is equal to zero, which means that these coefficients are de-correlated. The detail coefficients at different resolution levels are also de-correlated, which may simplify the statistical inference.

The second component of the compression module of a CS-based system is the measurement matrix $\mathbf{\Phi}$. As it was mentioned in Chapter 1, the universality property guarantees that a random measurement matrix is incoherent with an arbitrary transformation matrix $\mathbf{\Psi}$ with high probability. This means that the standard CS methods are non-adaptive, in the sense that the generation of CS measurements is based on an arbitrary random matrix and does not depend actually on the inherent statistical properties of the sparse signal. The fact that we are interested in achieving the highest possible sparsity of the signal in a suitable transform domain necessitates the use of statistical models that are able to support such an increased sparsity and adapt accordingly the measurement matrix so as to account for the specific statistical characteristics of the sparse signal.

Besides, the design of an efficient reconstruction Bayesian CS algorithm primarily depends on an accurate modeling of these characteristics. In the next chapter we review briefly the main properties of the family of α-Stable distributions, whose members are heavy-tailed and thus, suitable for representing highly sparse signals, as we will see in the subsequent analysis.

Chapter 3

Alpha-Stable models

As far as the laws of mathematics refer to reality they are not certain, and as far as they are certain they do not refer to reality.

A. Einstein (1879-1955)

3.1 Introduction

Figure 3.1 revisits the reconstruction module of a CS-based system. Given the CS measurements and the random measurement matrix $\mathbf{\Phi}$ the reconstruction algorithm recovers (in the general case) the sparse signal in the transform domain and then, the original signal is obtained simply by applying the inverse transform $\mathbf{\Psi}^{-1}$. As it will be clarified in the next chapters, the performance of a CS reconstruction algorithm when working in a Bayesian framework depends primarily on the selection of an accurate statistical model for the representation of the statistical behavior of the sparse signal.

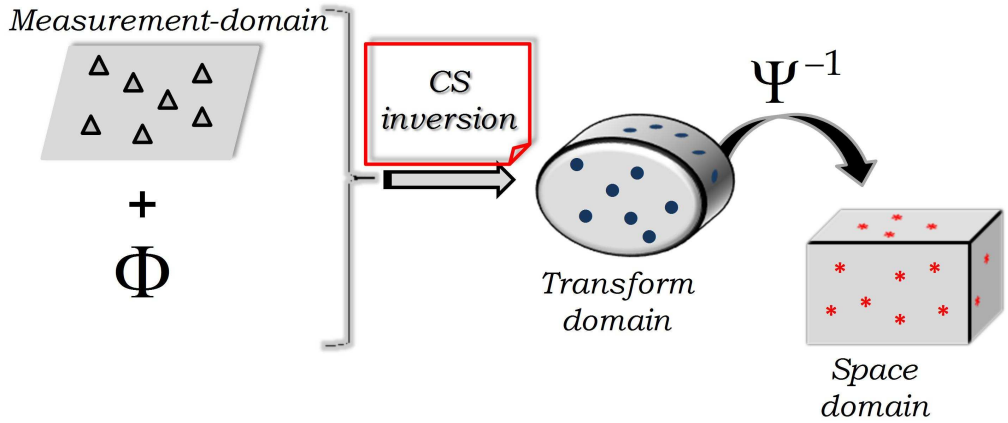


Figure 3.1: The decompression/reconstruction module of a CS-based system.

The majority of signal processing applications have been dominated traditionally by the Gaussian assumption, since the Gaussian model is justified by the Central Limit Theorem and also it often leads to analytical solutions. However, the statistical characteristics in a highly sparse scenario may be non-Gaussian. As a consequence, the Gaussian assumption may result in significant performance degradation for systems operating in a non-Gaussian environment. Phenomena that belong in the non-Gaussian class exhibit an impulsive nature. For instance,

a highly sparse signal can be considered to be impulsive in the sense that from coefficients of negligible amplitude it presents abrupt changes to coefficients of significant amplitude. This fact is expressed by the heavy-tailed behavior of their density functions, which means that the probability of large observations is non-negligible. As a result, the rate of decay in the tails of the distributions is less rapid than in the Gaussian case.

The heavy-tailed behavior of the distribution functions is described by a *stable law*, which is a direct generalization of the Gaussian distribution and includes the Gaussian as a limiting case. The class of *stable distributions* was first characterized by Lévy [100]. Since then, significant progress has been made in developing theory for stable processes [101, 102, 103, 104, 105, 106]. Applications of the stable distributions in the fields of signal processing and communications can be also found in [107, 108, 109, 110, 111]. This chapter introduces the basic concepts of a statistical model based on a sub-class of the stable family, namely, the class of *Symmetric-α-Stable* ($S\alpha S$) *distributions*.

3.2 The family of Stable Distributions

3.2.1 Univariate $S\alpha S$ Distributions

A symmetric α-Stable ($S\alpha S$) distribution is best defined by its characteristic function as follows:

$$\phi(t) = \exp(\imath\delta t - \gamma|t|^{\alpha}) \ , \tag{3.1}$$

where α is the *characteristic exponent*, taking values $0 < \alpha \leq 2$, δ ($-\infty < \delta < \infty$) is the *location parameter*, and γ ($\gamma > 0$) is the *dispersion* of the distribution. The characteristic exponent is a shape parameter which controls the "thickness" of the tails of the density function. The smaller the α is, the heavier the tails of the $S\alpha S$ density function. The dispersion parameter determines the spread of the distribution around its location parameter, similar to the variance of the Gaussian. When $1 < \alpha \leq 2$, the location parameter δ equals the mean of the $S\alpha S$ distribution, while for $0 < \alpha \leq 1$, δ corresponds to its median.

A $S\alpha S$ distribution is called *standard* if $\delta = 0$, $\gamma = 1$. For the parameterization of a $S\alpha S$ random variable X, corresponding to the characteristic function of Eq. (3.1), it holds that if X follows a $S\alpha S$ distribution with parameters α, γ, δ ($X \sim p_{\alpha}(\gamma, \delta)$), then the variable $(X - \delta)/\gamma^{1/\alpha}$ is standard with characteristic exponent α. Examples of standard $S\alpha S$ density functions for different values of α are shown in Figure 3.2 on the next page.

There are multiple parameterizations of the general univariate stable distributions which are useful for different problems. Two of them are the following [112]:

$$\phi(t) = \exp(\imath\delta t - \gamma^{\alpha}|t|^{\alpha}) \ , \tag{3.2}$$

$$\phi(t) = \exp(\imath\delta t - \frac{1}{\alpha}\,\gamma_{*}^{\alpha}|t|^{\alpha}) \ , \tag{3.3}$$

with $\gamma_{*} = \alpha^{1/\alpha}\gamma$. Parameterization (3.3) is probably the most intuitive for users in applied fields, as it has a number of interesting properties [112]: for instance, the mode of a $S\alpha S$ density following this parameterization is at δ and in the Gaussian case ($\alpha = 2$) γ_{*} is the standard deviation.

In general, no closed-form expressions exist for most density and distribution functions. Two important special cases of $S\alpha S$ densities with closed-form expressions are the Gaussian ($\alpha = 2$) and the Cauchy ($\alpha = 1$). Using parameterization (3.1) we have:

Gaussian

$$p_2(\gamma,\, \delta;\, x) = \frac{1}{\sqrt{4\pi\gamma}}\, e^{-\frac{(x-\delta)^2}{4\gamma}} \tag{3.4}$$

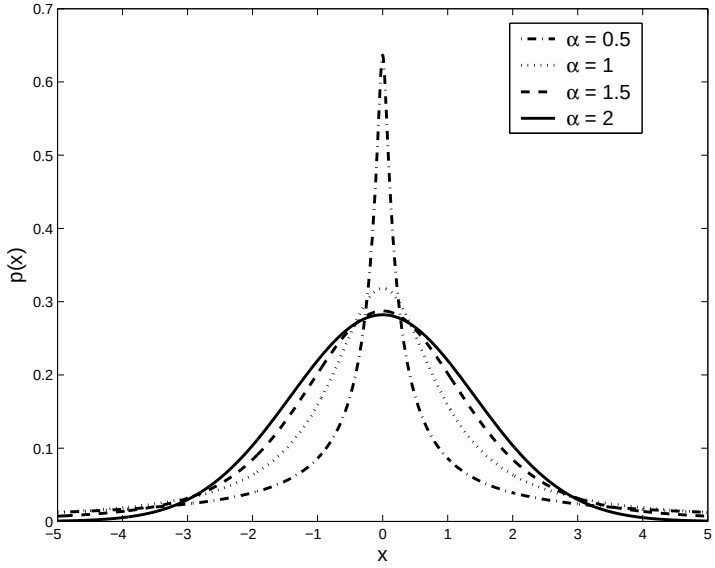


Figure 3.2: Standard $S\alpha S$ density functions for different values of the characteristic exponent α (parameterization (3.2)).

Cauchy

$$p_1(\gamma,\, \delta;\, x) = \frac{1}{\pi}\frac{\gamma}{\gamma^2 + (x-\delta)^2} \tag{3.5}$$

Note that when using parameterization (3.1) with $\alpha = 2$ (Gaussian) then the dispersion γ equals the half of the variance.

$S\alpha S$ densities have many common features with the Gaussian, they are smooth, unimodal, symmetric with respect to the median and bell-shaped. However, unlike the Gaussian density, which has exponential tails, the stable densities have tails following an algebraic rate of decay which makes them heavier than the Gaussian tails, as shown in Figure 3.3 on the following page. Figure 3.4 on page 31 displays some simulated $S\alpha S$ time series demonstrating the fact that random variables following $S\alpha S$ distributions with small α values are highly impulsive.

Two of the most important properties of the stable distributions are the *stability property* and the *Generalized Central Limit Theorem*, which are stated as follows:

Stability property: if $X_1, X_2, \ldots, X_N$ are independent random variables following stable distributions with the same α, then all linear combinations $\sum_{k=1}^{N} a_k X_k$, for arbitrary constants a_k, are stable with the same α.

Generalized Central Limit Theorem: X is stable if and only if it is the limit in distribution of sums of the form

$$\frac{X_1 + \cdots + X_N}{a_N} - b_N$$

where $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables and $\{a_N\}$, $\{b_N\}$ are sequences of positive and real numbers, respectively.

An important characteristic of the $S\alpha S$ distributions is the non-existence of the second-order moment, except for the Gaussian case ($\alpha = 2$). Instead, all moments of order q less than α do exist and are called the *Fractional Lower Order Moments* (FLOMs). The following theorem holds [102]:

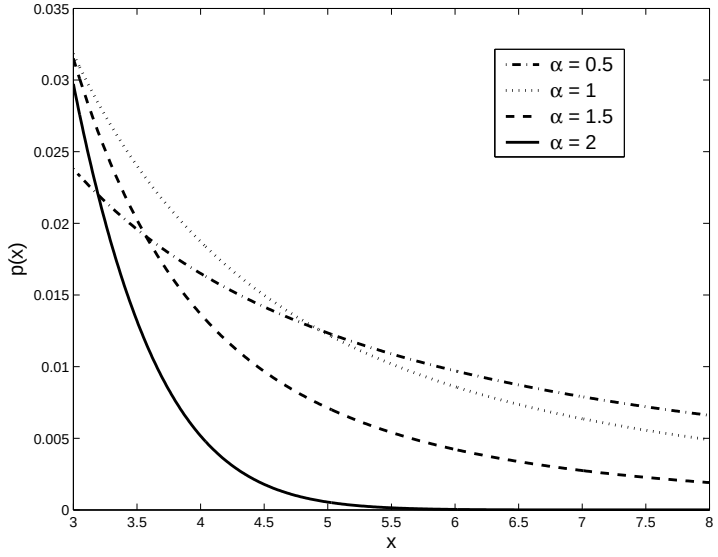


Figure 3.3: Tails of the standard $S\alpha S$ density functions for different values of the characteristic exponent α (parameterization (3.2)).

Theorem 3.1 (FLOM) *The FLOMs of a $S\alpha S$ random variable X, following parameterization* (3.2) *with zero location parameter and dispersion γ, are given by*

$$E\{|X|^q\} = \big(C(q,\alpha)\cdot\gamma\big)^q\,, \qquad for\ 0<q<\alpha \tag{3.6}$$

where

$$\big(C(q,\alpha)\big)^q = \frac{2^{q+1}\Gamma\Big(\frac{q+1}{2}\Big)\Gamma\Big(-\frac{q}{\alpha}\Big)}{\alpha\sqrt{\pi}\,\Gamma\Big(-\frac{q}{2}\Big)} = \frac{\Gamma\big(1-\frac{q}{\alpha}\big)}{\cos\big(\frac{\pi}{2}q\big)\Gamma(1-q)} \tag{3.7}$$

with $\Gamma(\cdot)$ denoting the Gamma function, which is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}\,e^{-t}\,\mathrm{d}t\ . \tag{3.8}$$

The practical implementation of a stable model is a non-trivial task due to the lack of analytical expressions for its probability density and distribution functions. DuMouchel [106] developed a (computationally intensive) procedure for approximating stable densities and distribution functions. Nolan [113] improved the approximation procedure by developing methods which make it possible to compute densities, distribution functions and quantiles for stable distributions quickly.

A very important task in designing algorithms based on the $S\alpha S$ models is the parameter estimation, that is, the estimation of the three parameters (α, γ, δ) describing the $S\alpha S$ distribution. Most of the conventional methods in mathematical statistics, such as the Maximum Likelihood (ML) estimation, can not be used in the stable case, since they depend on an explicit form for the density. Nevertheless, there are suboptimal numerical methods that have been found useful and efficient in practical applications. Among them, there are methods based on sample quantiles [114] and sample characteristic functions [115, 116]. There are also ML methods based on an approximation to the likelihood function [113, 117].

In the following chapters, we assume a $S\alpha S$ distribution located around the origin (i.e., $\delta = 0$) and we estimate the $S\alpha S$ model parameters (α, γ) using the consistent ML method

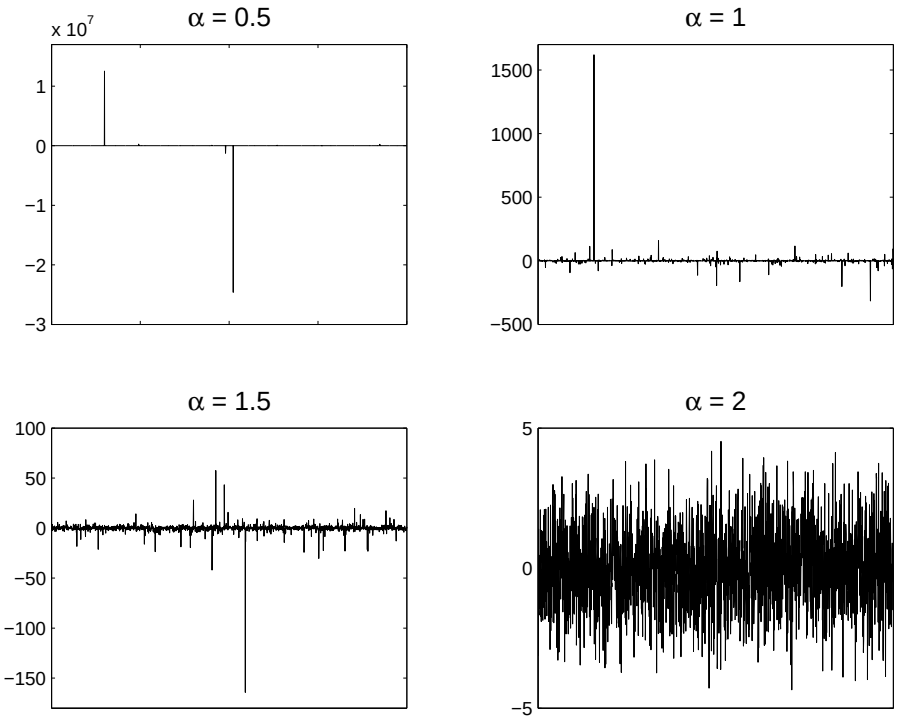


Figure 3.4: Simulated $S\alpha S$ sequences for different values of the characteristic exponent.

described by Nolan [113], which gives reliable estimates and provides the most tight confidence intervals.

3.2.2 Multivariate Stable Distributions

Unlike the family of univariate stable distributions, the family of multivariate stable distributions forms a non-parametric set (except for the multivariate Gaussian case ($\alpha = 2$)). The definition of the multivariate stable distribution is as follows:

Definition 3.1 (Multivariate stable distribution) *An N-dimensional distribution function $F(\vec{x})$, $\vec{x} \in \mathbb{R}^N$, is called stable if, for any i.i.d. random vectors $\vec{X}_1$, $\vec{X}_2$ with distribution function $F(\vec{x})$ and arbitrary constants a_1, a_2, there exist $a \in \mathbb{R}$, $\vec{b} \in \mathbb{R}^N$ and a random vector $\vec{X}$ with the same distribution function $F(\vec{x})$ such that*

$$a_1\vec{X}_1 + a_2\vec{X}_2 \stackrel{d}{=} a\vec{X} + \vec{b}$$

where $\stackrel{d}{=}$ means equality in distribution.

A multivariate stable distribution is determined by a location vector $\vec{\delta} \in \mathbb{R}^N$, the characteristic exponent $0 < \alpha \leq 2$ and a finite measure $\mu(\mathrm{d}\vec{s})$ on the unit sphere S_N of $\mathbb{R}^N$ [102]. In the multivariate case, a real random vector $\vec{X} = (X_1, \ldots, X_N)$ is $S\alpha S$, or the real random variables $X_1, \ldots, X_N$ are jointly $S\alpha S$, if their joint characteristic function is of the form

$$\phi(\vec{\theta}\,) = \exp\Big\{\imath\vec{\theta}^{\,T}\vec{\delta} - \int_{S_N} |\,\vec{\theta}^{\,T}\vec{s}\,|^{\alpha}\,\mu(\mathrm{d}\vec{s})\Big\} \tag{3.9}$$

with T denoting a transpose and where the spectral measure $\mu(\cdot)$ is symmetric, that is, $\mu(A) = \mu(-A)$ for any measurable set A on S_N.

A very important difference between the multivariate Gaussian and stable distributions is that a stable random vector can not be whitened. Specifically, it is known that if $\vec{X}$ is a Gaussian vector, then it can be written as

$$\vec{X} = \mathbf{A}\vec{Y}$$

where $\mathbf{A}$ is a constant matrix and $\vec{Y}$ is a Gaussian random vector with independent components. In the stable case, it is shown that representation of even two stable variables with characteristic exponent $0 < \alpha < 2$ as a linear combination of a finite number of independent stable variables of the same α is impossible [118].

Sub-Gaussian random vectors

An important sub-class of multivariate $S\alpha S$ random vectors is the class of the so-called α-*sub-Gaussian* random vectors [102] defined as follows,

Definition 3.2 (sub-Gaussian $S\alpha S$ random vector) *Any vector $\vec{X}$ distributed as*

$$\vec{X} = A^{1/2}\,\vec{G} \tag{3.10}$$

where A is a positive $\frac{\alpha}{2}$-stable random variable and $\vec{G} = (G_1, G_2, \ldots, G_N)$ is a zero-mean Gaussian random vector, independent of A, with covariance matrix $\boldsymbol{R}$, is called sub-Gaussian $S\alpha S$ random vector in $\mathbb{R}^N$ with underlying Gaussian vector $\vec{G}$.

This sub-class is often denoted by α-SG($\mathbf{R}$). Thus, a sub-Gaussian $S\alpha S$ random vector can be also viewed as a variance mixture of Gaussian vectors. An advantage of the sub-Gaussian $S\alpha S$ random vectors as a modeling tool is the simple analytical expression of the corresponding characteristic function. The following proposition holds:

Proposition 3.1 *The characteristic function of a sub-Gaussian $S\alpha S$ random vector $\vec{X}$ is given by,*

$$\begin{aligned}\phi(\vec{\theta}) &= \phi(\theta_1, \ldots, \theta_N) = \mathbb{E}\{e^{\imath\langle\vec{\theta},\vec{X}\rangle}\} \\ &= \mathbb{E}\{e^{\imath\sum_{k=1}^{N}\theta_k X_k}\} = \exp\Big\{-\Big|\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\theta_i\theta_j\boldsymbol{R}_{ij}\Big|^{\alpha/2}\Big\}\end{aligned} \tag{3.11}$$

where $\boldsymbol{R}_{ij} = \mathbb{E}\{G_iG_j\}$ are the covariances of the underlying Gaussian random vector $(G_1, \ldots, G_N)$,

with $\langle\cdot,\cdot\rangle$ denoting the inner product of two vectors.

We also note that in the zero-mean case the covariance of two random variables equals their correlation and thus, the $N \times N$ matrix $\mathbf{R}$ can be also viewed as the correlation matrix of the underlying vector $\vec{G}$. Eq. (3.11) can be rewritten as follows

$$\phi(\vec{\theta}) = \exp\Big\{-\Big|\frac{1}{2}\,\vec{\theta}^{\,T}\mathbf{R}\vec{\theta}\Big|^{\alpha/2}\Big\}$$

and taking into account that $\mathbf{R}$ is a non-negative definite matrix we can further simplify it as,

$$\phi(\vec{\theta}) = \exp\Big\{-\Big(\frac{1}{2}\,\vec{\theta}^{\,T}\mathbf{R}\vec{\theta}\Big)^{\alpha/2}\Big\}\ . \tag{3.12}$$

The above proposition results in the next corollary [102]:

Corollary 3.1 *Let $\vec{X}$ be sub-Gaussian $S\alpha S$ random vector with underlying Gaussian vector $\vec{G}$. Then there is a one-to-one correspondence between the probability distribution of $\vec{G}$ and that of $\vec{X}$.*

It is shown [102] that a vector with independent non-zero components cannot be sub-Gaussian. The components of a sub-Gaussian $S\alpha S$ random vector are, in fact, strongly dependent.

3.2.3 Covariations

The concept of covariance plays a fundamental role in the second-order moment theory. However, covariances do not exist in the family of $S\alpha S$ random variables due to their infinite variance. Instead, a quantity called *covariation*, which under certain constraints plays an analogous role for $S\alpha S$ random variables to the one played by covariance for Gaussian random variables, has been proposed:

Definition 3.3 (Covariation) *Let X and Y be jointly $S\alpha S$ random variables with $1 < \alpha \leq 2$, zero location parameters and dispersions γ_X and γ_Y respectively. Then for all $1 < q < \alpha$, the covariation of X with Y is defined by*

$$[X, Y]_\alpha = \frac{\mathbb{E}\{XY^{<q-1>}\}}{\mathbb{E}\{|Y|^q\}} \|Y\|_\alpha^\alpha \tag{3.13}$$

where for any complex number z and $a \geq 0$ we use the notation $z^{<a>} = |z|^{a-1}\bar{z}$ with $\bar{z}$ denoting complex conjugation,

where $\|Y\|_\alpha$ denotes the covariation norm of Y expressed in terms of its scale parameter as shown in Proposition 3.2 below. Linearity in the first argument and scaling are two useful properties of the covariation. The interested reader may refer to [102] for a more thorough analysis.
(Linearity in the first argument): If X_1, X_2, Y are jointly $S\alpha S$ then

$$[aX_1 + bX_2, Y]_\alpha = a[X_1, Y]_\alpha + b[X_2, Y]_\alpha \tag{3.14}$$

for any constants $a, b \in \mathbb{R}$.
(Pseudo-linearity in the second argument): If X_1, X_2, Y are jointly $S\alpha S$ then

$$[Y, aX_1 + bX_2]_\alpha = a^{<\alpha-1>}[Y, X_1]_\alpha + b^{<\alpha-1>}[Y, X_2]_\alpha \tag{3.15}$$

for any constants $a, b \in \mathbb{R}$.
(Scaling): If X, Y are jointly $S\alpha S$ and $a, b \in \mathbb{R}$, then

$$[aX, bY]_\alpha = ab^{<\alpha-1>}[X, Y]_\alpha \ . \tag{3.16}$$

The *covariation coefficient* of X with Y, for all $1 < q < \alpha$, is defined by:

$$\lambda_{XY} = \frac{[X, Y]_\alpha}{[Y, Y]_\alpha} = \frac{\mathbb{E}\{XY^{<q-1>}\}}{\mathbb{E}\{|Y|^q\}} \tag{3.17}$$

Note the asymmetric nature of the covariation and the covariation coefficient. However, we can define a *symmetric covariation coefficient* as follows:

$$Corr_\alpha(X, Y) = \lambda_{XY}\lambda_{YX} = \frac{[X, Y]_\alpha [Y, X]_\alpha}{[Y, Y]_\alpha [X, X]_\alpha} \ . \tag{3.18}$$

Let $\mathcal{S}_\alpha$ be the linear space of jointly $S\alpha S$ random variables. For $1 < \alpha \leq 2$, the covariation induces a norm on $\mathcal{S}_\alpha$,

Definition 3.4 (Covariation norm) *The covariation norm of $X \in \mathcal{S}_\alpha$, $1 < \alpha \leq 2$, is defined by*

$$\|X\|_\alpha = ([X, X]_\alpha)^{1/\alpha}$$

Depending on the parameterization of the characteristic function of a $S\alpha S$ distribution the following proposition holds:

Proposition 3.2 *If $X \sim p_\alpha(\gamma, 0)$ with $\alpha > 1$, then*

$$\|X\|_\alpha = \gamma^{1/\alpha} \ , \quad for\ parameterization\ (3.1)$$

and

$$\|X\|_\alpha = \gamma \ , \quad for\ parameterization\ (3.2)$$

Estimation of Covariations

In general, it is difficult to find analytical expressions for the covariations (or covariation coefficients). In practical applications it is important to have unbiased and efficient estimators of these quantities. The covariation coefficient λ_{XY} was defined as a fraction of the covariation $[X, Y]_\alpha$ and the scale parameter $[Y, Y]_\alpha$. Thus, having an estimation of the covariation coefficient λ_{XY} then, we can estimate $[X, Y]_\alpha$ by multiplying λ_{XY} and $[Y, Y]_\alpha$.

Focusing our attention on estimating the covariation coefficient λ_{XY} we present two methods that achieved an increased performance in terms of computational efficiency and estimation accuracy, by employing Monte-Carlo simulations. The first method, called the *Fractional Lower Order Moment (FLOM) Estimator* is very simple and computationally efficient and gives very good results especially when α is close to 2. The second method, called the *Screened Ratio (SR) Estimator*, exhibits a better performance for small values of α and is strongly consistent. In the following, we make the assumption of jointly $S\alpha S$ random variables X, Y with $\alpha > 1$ and a set of *independent* observations $(X_1, Y_1), \ldots, (X_N, Y_N)$ [119].

FLOM Estimator: For some $1 \leq q < \alpha$, the value of λ_{XY} is estimated by,

$$\hat{\lambda}_{FLOM} = \frac{\sum_{i=1}^{N} X_i |Y_i|^{q-1} \, sign(Y_i)}{\sum_{i=1}^{N} |Y_i|^q} \tag{3.19}$$

where

$$sign(x) = \begin{cases} 1 \ , & \text{for } x > 0 \\ 0 \ , & \text{for } x = 0 \\ -1 \ , & \text{for } x < 0 \end{cases} \tag{3.20}$$

Screened Ratio Estimator: For arbitrary constants c_1, c_2, with $0 < c_1 < c_2 \leq \infty$ the SR estimator of λ_{XY} is given by,

$$\hat{\lambda}_{SR} = \frac{\sum_{i=1}^{N} X_i Y_i^{-1} \chi_{Y_i}}{\sum_{i=1}^{N} \chi_{Y_i}} \tag{3.21}$$

with the random variable χ_Y defined as follows:

$$\chi_Y = \begin{cases} 1 \ , & \text{if } c_1 < |Y| < c_2 \\ 0 \ , & \text{otherwise.} \end{cases}$$

We usually choose $c_2 = \infty$ and c_1 to be a relatively small number.

The above estimators can be employed to estimate covariations between sub-Gaussian random variables or vectors. Consider the α-SG($\mathbf{R}$) vector $\vec{X} = A^{1/2} \vec{G}$. Then, the covariations

$\{[X_i, X_j]_\alpha\}_{i,j=1}^N$ are given by [102]:

$$[X_i, X_j]_\alpha = 2^{-\frac{\alpha}{2}} \, [\mathbf{R}]_{ij} \, [\mathbf{R}]_{jj}^{\frac{(\alpha-2)}{2}} \ . \tag{3.22}$$

Notice that if $[\mathbf{R}]_{ii} = [\mathbf{R}]_{jj}$ then the covariation is symmetric, that is, $[X_i, X_j]_\alpha = [X_j, X_i]_\alpha$. In the following, the covariations will be estimated using parameterization (3.2) for the characteristic function. By combining Eq. (3.13) with Proposition 3.2 we obtain,

$$[X, Y]_\alpha = \frac{\mathbb{E}\{XY^{<q-1>}\}}{\mathbb{E}\{|Y|^q\}} \, \gamma_Y^\alpha \ . \tag{3.23}$$

A covariation estimator is obtained by combining Eqs. (3.19), (3.23) as follows:

$$\hat{\mathbf{c}}_{ij}^{FLOM} = \frac{\sum_{k=1}^K \vec{X_i^k} \, |\vec{X_j^k}|^{q-1} \, sign(\vec{X_j^k})}{\sum_{k=1}^K |\vec{X_j^k}|^q} \, \gamma_{X_j}^\alpha \ , \tag{3.24}$$

where the dispersion γ_{X_j} is estimated using the ML estimator and the vectors $\{\vec{X^1}, \vec{X^2}, \ldots, \vec{X^K}\}$ constitute a set of K independent realizations of an α-SG($\mathbf{R}$) process, with $\vec{X^k} = (X_1^k, \ldots, X_N^k)$, $k = 1, \ldots, K$. In the stable case we define the *covariation matrix* $\mathbf{C}$ with its elements being the covariations. Then, the estimated covariation matrix $\hat{\mathbf{C}}$ is the matrix with elements $[\hat{\mathbf{C}}]_{ij} = \hat{\mathbf{c}}_{ij}^{FLOM}$.

Finally, from Eq. (3.22) we can estimate the elements $[\mathbf{R}]_{ij}$ of the underlying covariance matrix as follows:

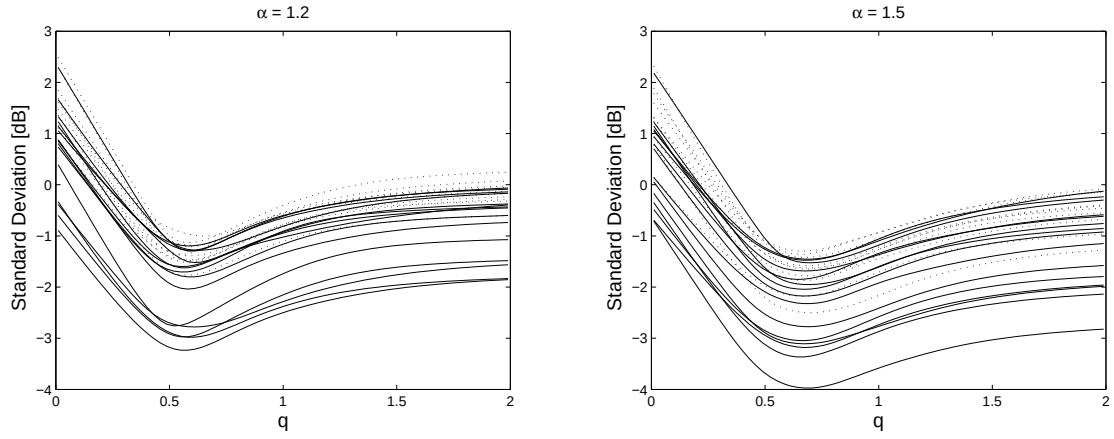
$$[\hat{\mathbf{R}}]_{jj} = \left(2^{\frac{\alpha}{2}} \, [\hat{\mathbf{C}}]_{jj}\right)^{\frac{2}{\alpha}}, \quad [\hat{\mathbf{R}}]_{ij} = 2^{\frac{\alpha}{2}} \, \frac{[\hat{\mathbf{C}}]_{ij}}{[\hat{\mathbf{R}}]_{jj}^{\frac{(\alpha-2)}{2}}} \tag{3.25}$$

which are consistent and asymptotically normal.

However, the estimation of covariations and covariation coefficients requires the specification of the arbitrary parameter q. When employing the multivariate sub-Gaussian model we need to estimate the covariations between the components of sub-Gaussian vectors. For this purpose, we repeat the process described in [120] and we create a table with the optimal values of q as a function of the characteristic exponent α, by finding the value of q that minimizes the standard deviation of the $\hat{\mathbf{c}}_{ij}^{FLOM}$ estimator, for different values of $\alpha > 1$. For this purpose we studied the influence of q on the performance of the covariation estimator via Monte-Carlo runs, using two sub-Gaussian random variables: $X = A^{1/2} G_X$, $Y = A^{1/2} G_Y$.

By definition, X and Y can be viewed as $S\alpha S$ random variables with dispersion γ_X and γ_Y, respectively. We generate a sample of a sub-Gaussian random variable by first generating a sample A drawn from an $S_{\alpha/2}((\cos\frac{\pi\alpha}{4})^{2/\alpha}, 1, 0)$ distribution [102] and then by generating a sample G drawn from a zero-mean Gaussian distribution with variance $2\gamma^2$, which is viewed as $S_2(\gamma, 0, 0)$ (with $\gamma = \gamma_X$ or $\gamma = \gamma_Y$ depending on whether the Gaussian part G corresponds to the variable X or Y, respectively). We executed $K = 1000$ Monte-Carlo runs with $\alpha \in [1 : 0.05 : 2]$ and for dispersions (γ_X, γ_Y) ranging in the interval $[0.01 : 0.1 : 5]$. Figure 3.5 displays the curves representing the standard deviation of the $\hat{\mathbf{c}}^{FLOM}(q)$ covariation estimator as a function of q for two values of α and 25 pairs of dispersions (γ_X, γ_Y). We observe that for each α all the curves are minimized in a common interval on the q-axis and actually their optimal value of q is close to each other. For a given α_i the optimal q_i is estimated by averaging the optimal q values over its corresponding curves and over the Monte-Carlo runs.

Table 3.1 shows the corresponding optimal values of q for several values of α. This table is also used as a lookup table in order to find the optimal q for every $1 < \alpha \leq 2$ by interpolating these values or by extrapolating them when the numerical ML estimation results in $\alpha < 1$.



(a) (b)

Figure 3.5: Curves representing the standard deviation of the $\hat{\mathbf{c}}^{FLOM}$ covariation estimator as a function of q for $\alpha = 1.2$, 1.5 and 25 dispersion pairs (γ_X, γ_Y).

α	Optimal q	α	Optimal q
1	0.52	1.5	0.69
1.05	0.54	1.55	0.71
1.1	0.56	1.6	0.72
1.15	0.57	1.65	0.74
1.2	0.58	1.7	0.76
1.25	0.59	1.75	0.79
1.3	0.61	1.8	0.81
1.35	0.62	1.85	0.84
1.4	0.64	1.9	0.88
1.45	0.66	1.95	0.93
		2	0.8

Table 3.1: Optimal q parameter as a function of the characteristic exponent α.

3.3 Statistical Modeling of Wavelet Subband Coefficients

The development of CS reconstruction methods in a transform-domain is based on the observation that often a linear, invertible transform re-structures the original signal, resulting in a set of transform coefficients whose structure is "simpler" to model. Real-world signals are characterized by a set of "features", such as abrupt changes for 1-D signals and edges, ridges and lines for images. For such signals, the wavelet transform is a powerful modeling tool as it was mentioned in Chapter 2. The following properties of the wavelet transform justify the design and implementation of CS methods in the wavelet domain:

❶ *Locality*: Each wavelet coefficient represents signal content localized in both space and frequency.

❷ *Multiresolution*: The wavelet transform decomposes each signal at a nested set of scales.

❸ *Abrupt change detection*: Wavelet functions operate as local abrupt change detectors, representing them by large wavelet coefficients at the corresponding locations.

Consequently, we can closely approximate a signal using just a few, large amplitude, wavelet coefficients.

The development of efficient CS reconstruction algorithms working in a Bayesian framework require the accurate modeling of the marginal densities of wavelet subband coefficients. These coefficients have been commonly represented as Gaussian random variables, since this assumption results in convenient expressions. However, the above mentioned properties of the wavelet transform yield that the set of wavelet coefficients of real-world signals tends to be sparse, resulting in a large number of small amplitude coefficients and a small number of large amplitude coefficients. This property is in conflict with the Gaussian assumption and thus, may degrade the performance of a CS-based system giving rise to peaky and heavy-tailed *non-Gaussian* marginal distributions of the wavelet subband coefficients. This also justifies the use of $S\alpha S$ distributions in the design of Bayesian CS reconstruction techniques.

Closing this chapter, we provide examples which show that an often better approximation of the marginal density of coefficients at distinct subbands, produced by various types of wavelet transforms, may be obtained by alpha-Stable models. The $S\alpha S$ model is suitable for describing signals with heavier distribution tails than what is assumed by exponential families, like the Gaussian and the *generalized Gaussian density* (GGD). Indeed, the $S\alpha S$ density follows an *algebraic rate* of decay that depends on the value of the characteristic exponent: $P\{X > x\} \sim c_\alpha x^{-\alpha}$.

The statistical fitting proceeds in two steps: first, we assess whether the data deviate from the normal distribution and we determine if they have heavy tails by employing normal probability plots [121]. Then, we check if the data is in the stable domain of attraction by estimating the characteristic exponent α directly from the data and by providing the related confidence intervals. As further stability diagnostics, we employ the amplitude probability density (APD) curves ($P\{|X| > x\}$) that give a good indication of whether the $S\alpha S$ fit matches the data near the mode and on the tails of the distribution.

In the first illustration, the accuracy of the $S\alpha S$ model is tested on a 1-D underwater acoustics' signal, which is decomposed in 3 levels using Daubechies' 4 (db4) wavelet. Figure 3.6 on the following page shows the signal along with the APD curves corresponding to the empirical, the Gaussian, the GGD and the $S\alpha S$ distributions, at each decomposition level. It is clear that the $S\alpha S$ model provides a superior approximation near the mode and on the tails of the empirical distribution, when compared with the other models.

In the second example, the accuracy of the $S\alpha S$ model is tested on the 2-D "Indor 4" image, which is decomposed in 3 levels using the db4 wavelet. Figure 3.7 shows the image along with the APD curves corresponding to the empirical, the Gaussian, the GGD and the $S\alpha S$ distributions, for one direction per decomposition level suggestively (H: horizontal, D: diagonal, V: vertical). Although for the horizontal direction at the third level the $S\alpha S$ and GGD models results in comparable approximation capability, however, it is clear that the $S\alpha S$ model provides a superior approximation near the mode and on the tails of the empirical distribution for the rest of directions when compared with the other models. This behavior is valid for the other directions, too, which are not shown in the figure. The value of the characteristic exponent α is also included in parentheses in the title of each subplot. The fact that its value deviates largely from 2 indicates a highly impulsive nature of the corresponding wavelet coefficients. This is equivalent to a very large number of negligible-amplitude coefficients and only a few significant ones, which results in a highly sparse coefficient vector. This also justifies the appropriateness of alpha-Stable models in the design of efficient Bayesian CS methods.

In the following chapters, we introduce and describe in detail the design and implementation of novel Bayesian CS reconstruction methods, which exploit the sparsity-enforcing property of the wavelet transform, as well as the accurate modeling of the high sparsity behavior via members of the alpha-Stable family of distributions.

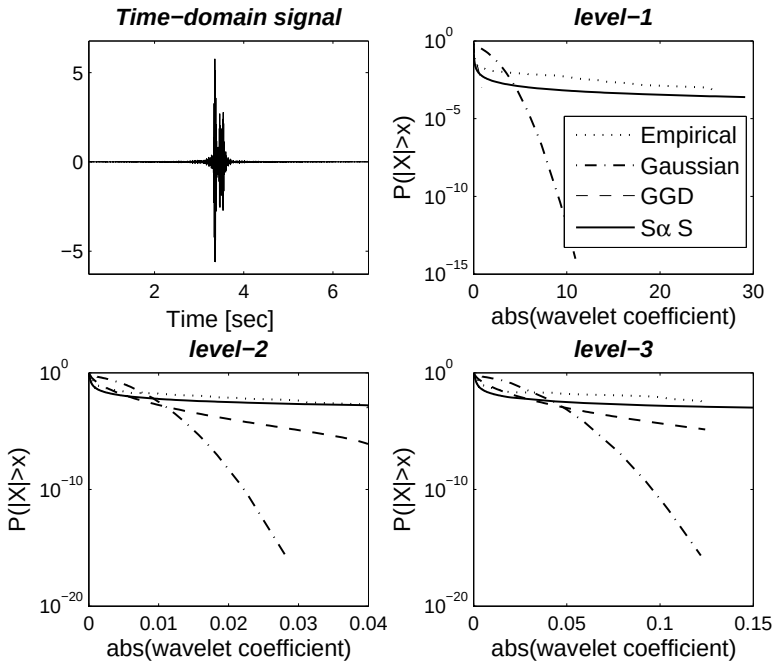


Figure 3.6: Acoustic signal and APD curves corresponding to the empirical, Gaussian, GGD and $S\alpha S$ distributions, at each decomposition level (3-level DWT (db4)).

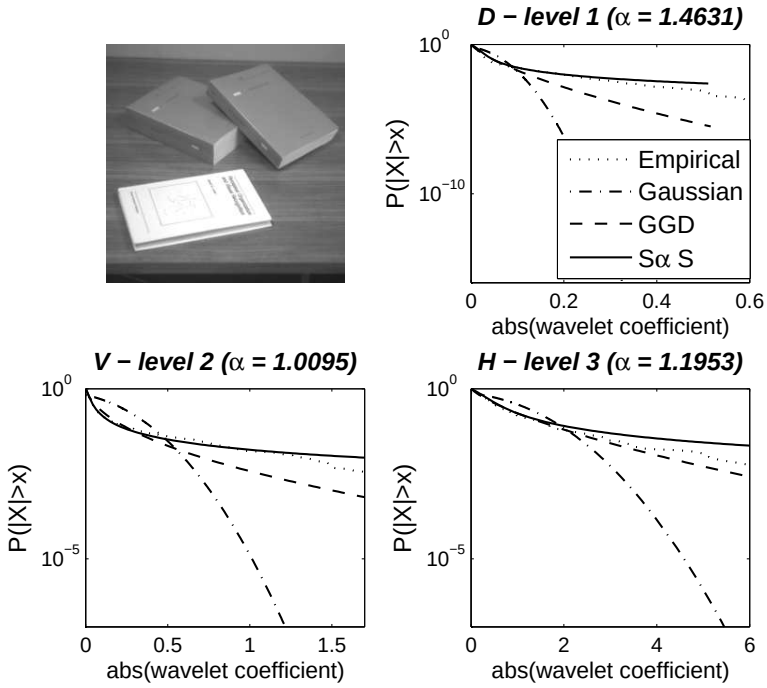


Figure 3.7: "Indor 4" image and APD curves corresponding to the empirical, Gaussian, GGD and $S\alpha S$ distributions, at each decomposition level (3-level DWT (db4)).

Part II

Heavy-tailed Bayesian Compressed Sensing Algorithms

Chapter

4

Bayesian Compressed Sensing via Gaussian Scale Mixtures

> Where is the knowledge we have lost in information?
>
> T. S. Eliot
> *The Rock (1934)*

4.1 Introduction

As it was mentioned in Chapter 1, sampling is a key concept of signal processing since it allows real-world signals in the continuous-domain to be acquired, represented and processed in the discrete-domain. The long-term trend of applying the classical Shannon/Nyquist sampling theorem specifies that to avoid losing information when capturing a signal we must sample at least twice faster than the signal bandwidth. In many modern applications, including digital image and video cameras, this sampling rate is so high resulting in too many samples, and thus, making compression a necessity prior to storage or transmission. In addition, there are cases where even higher compression rates would suffice to carry out a specific task, such as in image classification and retrieval, where a high-quality reconstruction of the still images is not necessary. In contrast, a highly sparse representation of the database images is required under the possible limited storage and power constraints of the imaging system.

Besides, in Chapter 2 we saw that many natural signals result in a highly compact (sparse) representation, when they are projected on localized orthonormal basis functions, such as the wavelets. The traditional approach to compressing such a sparse signal is to compute its transform coefficients and then store or transmit only a small number of large amplitude coefficients. However, this is an inherently wasteful process in terms of both sampling rate and computational complexity, since one gathers and processes the entire signal even though an exact representation is not required explicitly.

On the other hand, *compressed sensing* (CS) enables a potentially significant reduction in the sampling and computation costs at a sensing system with limited capabilities. According to the CS framework a signal having a sparse representation in a transform basis $\mathbf{\Psi}$ can be reconstructed from a small set of projections onto a second, measurement basis $\mathbf{\Phi}$ that is incoherent with the first one. Also recall that CS provides a simple compression scheme with low computational complexity, which is *asymmetrical* in nature with the compression part being of very low complexity (simple linear projections), while the main computational cost is on the decompression part where increased computational resources are available. This inherent property of CS enables its use in practical application involving digital devices with limited computational and power resources, such as in a wireless sensor network or in a network of cameras acquiring high-resolution images.

On the other hand, the majority of the works presented in the recent literature about the sparse representation and reconstruction of a signal in an over-complete dictionary using CS has concentrated on solving constrained-based optimization problems (cf. Section 1.2). For instance, following the ℓ_1-norm minimization approach in the general case of noisy measurements we seek the sparsest weight (coefficient) vector, which solves the problem (1.16). These approaches for the solution of such an optimization problem result in a *point estimate* of the weight vector $\vec{w}$.

In a recent study [55], the inversion of CS measurements was considered from a Bayesian perspective (Bayesian Compressed Sensing (BCS)). In particular, given a prior belief that the weight vector $\vec{w}$ should be sparse in the basis $\mathbf{\Psi}$ and the set of CS measurements $\vec{g}$ (observables), the objective is to formulate a *posterior probability distribution* for $\vec{w}$. This improved the accuracy over the point estimate and provided confidence intervals (error bars) in the approximation of the original signal $\vec{x}$. Besides, this was also used to guide the optimal design of additional CS measurements implemented with the goal of reducing the uncertainty in reconstructing $\vec{x}$. For computational purposes and in order to get closed-form expressions for the Bayesian inference, the prior distribution of $\vec{w}$ was approximated using a sparsity-enforcing hierarchical model.

In the present chapter, the CS-based estimation of $\vec{w}$ is also performed in a Bayesian framework. However, in contrast to the previous work, our proposed method consists in modeling directly the prior distribution of $\vec{w}$ with a heavy-tailed distribution, which promotes the sparsity of $\vec{w}$. This is motivated by the fact that a heavy-tailed density function is suitable for modeling highly impulsive signals. In our case, $\vec{w}$ could be considered as a highly impulsive "signal", since it is characterized by a large number of close to zero-amplitude components and a small number of large-amplitude components. For this purpose, a Gaussian Scale Mixture (GSM) is employed to model the heavy-tailed behavior of $\vec{w}$.

4.2 Bayesian CS inversion

Let $\mathbf{\Psi}$ be a $N \times N$ matrix, whose columns correspond to the transform basis functions. Then, a given signal $\vec{x} \in \mathbb{R}^N$ can be represented as $\vec{x} = \mathbf{\Psi}\vec{w}$, where $\vec{w} \in \mathbb{R}^N$ is the weight (coefficients) vector. As mentioned in Chapter 2, for many natural signals the majority of the components of $\vec{w}$ have negligible amplitude. In particular, $\vec{x}$ is L-sparse in basis $\mathbf{\Psi}$ if the corresponding weight vector $\vec{w}$ has L non-zero components ($L \ll N$). In a real-world scenario $\vec{x}$ is not strictly L-sparse, but it is said to be *compressible* when the re-ordered components of $\vec{w}$ decay at a power-law.

Also let $\mathbf{\Phi}$ be a $M \times N$ ($M < N$) measurement matrix as described in Section 1.1.1. For instance, let $\mathbf{\Phi}$ be a Hadamard matrix or a matrix containing independent and identically distributed (i.i.d.) Gaussian entries. Such matrices are incoherent with any fixed transform matrix $\mathbf{\Psi}$ with high probability (universality property) [4].

If $\vec{x}$ is compressible in $\mathbf{\Psi}$, then, it is possible to perform directly a compressed set of measurements $\vec{g}$ resulting in a simplified acquisition system. The original signal $\vec{x}$ and the CS measurements $\vec{g}$ are related through random projections, $\vec{g} = \mathbf{\Phi}\mathbf{\Psi}^T\vec{x} = \mathbf{\Phi}\vec{w}$, where $\mathbf{\Phi} = [\vec{\phi}_1, \ldots, \vec{\phi}_M]^T$ and $\vec{\phi}_m \in \mathbb{R}^N$ is a random vector with i.i.d. components. Thus, finding a sparse representation of $\vec{x}$ from $\vec{g}$ reduces to estimating a weight vector $\vec{w}$ with as few non-zero components as possible, which then can be used for reconstructing $\vec{x}$.

Notice that the representation of the original space-domain signal $\vec{x}$ is equivalent to its frequency-domain representation $\vec{w}$ (for instance, $\vec{w}$ may contain the wavelet coefficients of $\vec{x}$) related via the invertible matrix $\mathbf{\Psi}$. Thus, without loss of generality, in the following study we consider that the noisy CS measurements corresponding to the L-sparse signal $\vec{x}$ are acquired

in the transform domain using the model:

$$\vec{g} = \mathbf{\Phi}\vec{w} + \vec{\eta} \,, \tag{4.1}$$

where $\vec{\eta}$ is the associated additive noise component.

Most of the recent literature on CS has concentrated on solving constrained optimization problems for sparse signal representation. For instance, when the CS measurements are acquired via Eq. (4.1) an ℓ_1-norm minimization approach seeks a sparse vector $\vec{w}$ by solving the following optimization problem,

$$\tilde{\vec{w}} = arg\min_{\vec{w}} \|\vec{w}\|_1 \;,\; s.t. \;\; \|\vec{g} - \mathbf{\Phi}\vec{w}\|_\infty \leq \epsilon \;, \tag{4.2}$$

where ϵ is the noise level ($\|\vec{\eta}\|_2 \leq \epsilon$). The main approaches for the solution of such optimization problems include linear programming and greedy algorithms (cf. Section 1.2) resulting in a *point estimate* of the weight vector $\vec{w}$.

On the other hand, when the CS inversion is treated from a Bayesian perspective, then given a prior belief that $\vec{w}$ is sparse in basis $\mathbf{\Psi}$ and the set of CS measurements $\vec{g}$, the objective is to formulate a *posterior probability distribution* for $\vec{w}$. This improves the accuracy over a point estimate and provides confidence intervals (error bars) in the approximation of $\vec{x}$, which can be used to guide the optimal design of additional CS measurements with the goal of reducing the uncertainty in reconstructing $\vec{x}$.

By considering that the CS measurements are corrupted by additive, zero-mean Gaussian noise $\vec{\eta}$ with unknown variance σ_η^2 we obtain the following Gaussian likelihood model:

$$p(\vec{g}|\vec{w},\, \sigma_\eta^2) = (2\pi\sigma_\eta^2)^{-M/2} \cdot \exp\Big(-\frac{1}{2\sigma_\eta^2}\|\vec{g} - \mathbf{\Phi}\vec{w}\|\Big) \;. \tag{4.3}$$

Given the CS measurements $\vec{g}$ and assuming that the measurement matrix $\mathbf{\Phi}$ is known, the quantities to be estimated are the sparse weight vector $\vec{w}$ and the noise variance σ_η^2. Working in a Bayesian framework, this is equivalent to seeking a full posterior density function for $\vec{w}$ and σ_η^2.

In this probabilistic framework, the assumption that $\vec{w}$ is sparse is formalized by modeling the distribution of $\vec{w}$ using a sparsity-enforcing prior distribution. A common choice of this prior is the Laplace density [58]. Thus, given the CS measurements $\vec{g}$ and assuming the Gaussian likelihood model of Eq. (4.3), it is straightforward to see that the solution of the constrained optimization problem of Eq. (1.16) corresponds to a *maximum a posteriori* (MAP) estimate of $\vec{w}$ using an appropriate prior density.

4.3 Estimation of a sparse vector $\vec{w}$ using a GSM

The use of a Laplace prior density raised the problem that the Bayesian inference may not be performed in closed form [58], since the Laplace prior is not conjugate[1] to the Gaussian likelihood model. The treatment of the CS measurements $\vec{g}$ from a Bayesian viewpoint, while overcoming the problem of conjugateness, was introduced in [55]. In particular, rather than modeling the distribution of $\vec{w}$ using a Laplace prior, a hierarchical prior model was invoked using a set of hyperparameters, which had similar properties as the Laplace prior but allowed convenient conjugate-exponential analysis. Then, the overall prior on $\vec{w}$ was evaluated analytically resulting in the Student-t distribution [50], which can be considered as a sparseness prior, since it is peaked about zero.

[1] In probability theory, a family of prior probability distributions $p(s)$ is said to be conjugate to a family of likelihood functions $p(x|s)$ if the resulting posterior distribution $p(s|x)$ is in the same family as $p(s)$.

4.3.1 GSM prior model

In contrast to the previous BCS work, our proposed method consists in modeling directly the prior distribution of $\vec{w}$ using a heavy-tailed distribution, which promotes its sparsity, since it is suitable for modeling highly impulsive signals. This is motivated by the fact that the content of many natural signals (e.g., images) is often well structured (e.g., in the case of images containing edges) and thus $\vec{w}$ can be considered as highly impulsive, since it is characterized by a large number of negligible-amplitude components and a small number of large-amplitude components.

For this purpose, in our proposed method we replace the approximate hierarchical process by modeling directly the prior distribution of $\vec{w}$ by means of a Gaussian Scale Mixture (GSM).

Definition 4.1 (Gaussian Scale Mixture) *A vector $\vec{w} \in \mathbb{R}^N$ is called a GSM with underlying Gaussian vector $\vec{G}$ iff it can be written in the form $\vec{w} = A^{1/2}\,\vec{G}$, where A is a positive random variable and $\vec{G} = (G_1, G_2, \ldots, G_N)$ is a zero-mean Gaussian random vector, independent of A, with covariance matrix $\mathbf{\Sigma}$.*

Notice that when A is a positive $\frac{\alpha}{2}$-stable random variable the GSM model reduces to an α-SG($\mathbf{R}$) model (Definition 3.2). However, in the subsequent analysis we only consider that the components of $\vec{G}$ are also independent, and thus the covariance matrix becomes diagonal $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots \sigma_N^2)$, without making any assumption about the prior distribution of A.

From the above definition, the density of $\vec{w}$ conditioned on the variable A is a zero-mean multivariate Gaussian given by:

$$p(\vec{w}|A) = \frac{\exp\left(-\frac{1}{2}\vec{w}^T(A\mathbf{\Sigma})^{-1}\vec{w}\right)}{(2\pi)^{N/2}|A\mathbf{\Sigma}|^{1/2}}\ , \tag{4.4}$$

where $|\cdot|$ denotes the determinant of a matrix. From Eq. (4.4), we obtain the following simple expression for the maximum likelihood estimate of the variable A,

$$\hat{A}(\vec{w}) = \frac{\vec{w}^T\mathbf{\Sigma}^{-1}\vec{w}}{N}\ . \tag{4.5}$$

4.3.2 BCS inversion using a GSM

Assuming that the noise variance σ_η^2, the value of A, and the covariance matrix $\mathbf{\Sigma}$ have been estimated, given the CS measurements $\vec{g}$ and the random measurement matrix $\mathbf{\Phi}$, then, the posterior density of $\vec{w}$ is given by the Bayes' rule combining the likelihood and the prior density functions and exploiting the independence of A and $\vec{G}$,

$$p(\vec{w}|\vec{g}, A, \mathbf{\Sigma}, \sigma_\eta^2) = \frac{p(\vec{g}|\vec{w}, \sigma_\eta^2)p(\vec{w}|A, \mathbf{\Sigma})}{p(\vec{g}|A, \mathbf{\Sigma}, \sigma_\eta^2)}\ , \tag{4.6}$$

which results after some algebraic manipulation in a multivariate Gaussian distribution whose mean $\vec{\mu}$ and covariance $\mathbf{P}$ are given by

$$\vec{\mu} = \sigma_\eta^{-2}\mathbf{P}\mathbf{\Phi}^T\vec{g}\ , \tag{4.7}$$

$$\mathbf{P} = (\sigma_\eta^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{D})^{-1}\ , \tag{4.8}$$

where $\mathbf{D} = diag((A\sigma_1^2)^{-1}, \ldots, (A\sigma_N^2)^{-1})$. Working in this framework, the estimated weight vector $\vec{w}$ is chosen to be equal to the most probable value of the above multivariate Gaussian model, that is, $\vec{w} \equiv \vec{\mu}$.

Thus, the problem of estimating the sparse weight vector $\vec{w}$ reduces to estimating the unknown model parameters A, $\mathbf{\Sigma}$, σ_η^2. This is carried out by performing a *type-II* maximum

likelihood estimation. By noting that Eq. (4.8) can be re-written in the following equivalent form,

$$A^{-1}\mathbf{P} = \overbrace{\left(A\sigma_\eta^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{\Sigma}^{-1}\right)^{-1}}^{\tilde{\mathbf{P}}} , \tag{4.9}$$

we can estimate the unknown parameters σ_η^2, $\{\sigma_i^{-2}\}_{i=1}^N$ iteratively by maximizing the following marginal log-likelihood function based on the matrix $\tilde{\mathbf{P}}$ with respect to the unknown parameters:

$$\mathcal{L}(\sigma_\eta^2,\, \{\sigma_i^{-2}\}_{i=1}^N) = \log\Big[p(\vec{g}|A,\, \sigma_\eta^2,\, \{\sigma_i^{-2}\}_{i=1}^N)\Big] = -\frac{1}{2}\Big[M\, \log(2\pi) + \log(|\mathbf{C}|) + \vec{g}^T\mathbf{C}^{-1}\vec{g}\Big] , \tag{4.10}$$

where $\mathbf{C} = \frac{\sigma_\eta^2}{A}\mathbf{I} + \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T$. By comparing Eq. (4.10) with the marginal likelihood model in [59] we notice that our proposed model is actually a scaled version of the previous hierarchical models by a factor of A^{-1}. This factor plays an important role in the estimation process, since it controls the heavy-tailed behavior of the diagonal elements of $\mathbf{D}$ and consequently of the covariance matrix $\mathbf{P}$, and thus the sparsity of the estimated weight vector $\vec{w} \equiv \vec{\mu}$.

The fast update algorithm described in [59] is adapted and used for the maximization of Eq. (4.10) resulting in a sequential addition and deletion of candidate basis functions (columns of $\mathbf{\Phi}$)[2] to increase monotonically the marginal likelihood. First of all, note that the matrix $\mathbf{C}$ can be decomposed as follows:

$$\mathbf{C} = \frac{\sigma_\eta^2}{A}\mathbf{I} + \sum_{n=1, n\neq i}^{N} \sigma_n^2 \vec{\phi}_{\cdot,n}\vec{\phi}_{\cdot,n}^T + \sigma_i^2\vec{\phi}_{\cdot,i}\vec{\phi}_{\cdot,i}^T = \mathbf{C}_{-i} + \sigma_i^2\vec{\phi}_{\cdot,i}\vec{\phi}_{\cdot,i}^T , \tag{4.11}$$

where $\mathbf{C}_{-i}$ is $\mathbf{C}$ with the contribution of the i-th basis vector ignored. Thus, the determinant and the inverse of $\mathbf{C}$ can be computed sequentially using the following expressions:

$$|\mathbf{C}\,| = |\mathbf{C}_{-i}|\,|1 + \sigma_i^2\vec{\phi}_{\cdot,i}^T\mathbf{C}_{-i}^{-1}\vec{\phi}_{\cdot,i}| \tag{4.12}$$

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\vec{\phi}_{\cdot,i}\vec{\phi}_{\cdot,i}^T\mathbf{C}_{-i}^{-1}}{\sigma_i^{-2} + \vec{\phi}_{\cdot,i}^T\mathbf{C}_{-i}^{-1}\vec{\phi}_{\cdot,i}} . \tag{4.13}$$

Consequently, the marginal log-likelihood can be decomposed in two terms,

$$\mathcal{L}(\sigma_\eta^2,\, \{\sigma_i^{-2}\}_{i=1}^N) = \mathcal{L}(\sigma_\eta^2,\, \{\sigma_i^{-2}\}_{i=1, i\neq i'}^N) + l(\sigma_{i'}^{-2}) , \tag{4.14}$$

with the first term depending on all except for the i'-th variance, while the second term depends only on the i'-th variance. In particular, $l(\sigma_{i'}^{-2})$ is given by:

$$l(\sigma_{i'}^{-2}) = \frac{1}{2}\Big[\log(\sigma_{i'}^{-2}) - \log(\sigma_{i'}^{-2} + s_{i'}) + \frac{q_{i'}^2}{\sigma_{i'}^{-2} + s_{i'}}\Big] , \tag{4.15}$$

where $s_{i'} = \vec{\phi}_{\cdot,i'}^T\mathbf{C}_{-i'}^{-1}\vec{\phi}_{\cdot,i'}$ and $q_{i'} = \vec{\phi}_{\cdot,i'}^T\mathbf{C}_{-i'}^{-1}\vec{g}$. In order to monotonically maximize the marginal log-likelihood it suffices to focus on maximizing the second term of Eq. (4.14). It can be easily seen by analyzing Eq. (4.15) that it has a single maximum

$$\sigma_{i'}^{-2} = \begin{cases} \frac{s_{i'}^2}{q_{i'}^2 - s_{i'}} , & \text{if } q_{i'}^2 > s_{i'} , \\ \infty , & \text{if } q_{i'}^2 \leq s_{i'} . \end{cases} \tag{4.16}$$

[2]For convenience we denote the i-th column of $\mathbf{\Phi}$ by $\vec{\phi}_{\cdot,i}$, to avoid any confusion with its definition in Section 1.1.1. However, elsewhere in the text the single-subscript notation $\vec{\varphi}_i$ denotes a column vector.

This result implies that,

- ❒ if the i'-th basis function $\vec{\phi}_{\cdot,i'}$ has been included in the set of "significant" basis functions (that is, $\sigma_{i'}^{-2} < \infty$) in a previous iteration, while the current iteration yields $q_{i'}^2 \leq s_{i'}$, then it can be deleted from the model (that is, set $\sigma_{i'}^{-2} = \infty$).

- ❒ if the i'-th basis function $\vec{\phi}_{\cdot,i'}$ has not been included yet in the set of "significant" basis functions (that is, $\sigma_{i'}^{-2} = \infty$), while the current iteration yields $q_{i'}^2 > s_{i'}$, then it can be added in the model (that is, set $\sigma_{i'}^{-2}$ to a finite optimal value).

- ❒ if the i'-th basis function $\vec{\phi}_{\cdot,i'}$ has been already included in the set of "significant" basis functions (that is, $\sigma_{i'}^{-2} < \infty$) and the current iteration still yields $q_{i'}^2 > s_{i'}$, then the corresponding value of $\sigma_{i'}^{-2}$ can be re-estimated.

Note that, apart from an estimation of the optimal sparse vector $\vec{w}$, the algorithm also returns the set of significant basis functions denoted by $\mathscr{B}$. Besides, a very important difference of a BCS approach when compared with a norm-based one is that it permits successive additions and deletions of basis functions in the model. Thus, a basis function which may be considered to be significant at the beginning can be deleted in a subsequent iteration if the algorithm decides that it was not representative for the CS measurements. On the other hand, the norm-based CS reconstruction methods start with a single basis function and they can only add a new basis function in the model when moving from one iteration to the next one. This is the reason why BCS approaches often achieve an increased reconstruction performance, but at the cost of a possibly increased number of iterations.

Several convergence criteria can be employed to terminate the execution of the algorithm, such as when the number of iterations exceeds a predefined maximum or when the relative decrease of the marginal log-likelihood function from one iteration to the next one falls below a small positive threshold. In our implementation we adopt the second approach, since it results in an increased reconstruction performance, while the first one could be used to reduce the computational cost. The iterative scheme for the estimation of the weight vector $\vec{w}$ proceeds as shown by Algorithm 1.

Besides, it is important to note that the computational complexity of our proposed method, compared to the complexity of the standard BCS approach [55] is only slightly increased at the order of $\mathcal{O}(L^2)$ due to the estimation of A. This increase is negligible for truly sparse signals, since $L \ll N$.

4.4 Adaptive BCS using a GSM

As it was mentioned before, one of the advantages when working in a probabilistic framework is that we are able to provide a measure of uncertainty in the estimation of the original signal $\vec{x}$. This can be further employed to design adaptively the measurement matrix $\mathbf{\Phi}$ by selecting the next projection vector $\vec{\phi}_{M+1}$ with the goal of reducing the uncertainty of $\vec{x}$.

Since $\vec{x} = \mathbf{\Psi}\vec{w}$ and $\vec{w}$ follows a multivariate Gaussian density with a mean vector and a covariance matrix which are given by Eqs. (4.7), (4.8), respectively, we conclude that $\vec{x}$ also follows a multivariate Gaussian density with the following mean and covariance:

$$\mathbb{E}\{\vec{x}\} = \mathbf{\Psi}\vec{\mu}\,, \tag{4.17}$$

$$Cov\{\vec{x}\} = \mathbf{\Psi}\mathbf{P}\mathbf{\Psi}^T\,. \tag{4.18}$$

The diagonal elements of the above covariance matrix correspond to the error bars on the accuracy of the estimate of $\vec{x}$.

Algorithm 1 Estimation of the sparse weight vector $\vec{w}$ via BCS-GSM

Input: $\mathbf{\Phi}$, $\vec{g}$, $c \sim 10^{-3}$

Output: $\hat{\vec{w}} \equiv \vec{\mu}$, $\mathbf{P}$, σ_η^2, $\mathscr{B}$ {the set of significant basis functions}

Initialize: $\sigma_\eta^2 = c \cdot Var(\vec{g})$

select basis vector $\vec{\phi}_{\cdot,i_1}$ (i_1-th column of $\mathbf{\Phi}$) s.t. $i_1 = \arg\max_{i=1,\ldots,N} \frac{\|\vec{\phi}_{\cdot,i}\|^2}{\left(\|\vec{\phi}_{\cdot,i}^T \vec{g}\|^2 / \|\vec{\phi}_{\cdot,i}\|^2\right) - \sigma_\eta^2}$

set $\sigma_{i_1}^{-2} = \frac{\|\vec{\phi}_{\cdot,i_1}\|^2}{\left(\|\vec{\phi}_{\cdot,i_1}^T \vec{g}\|^2 / \|\vec{\phi}_{\cdot,i_1}\|^2\right) - \sigma_\eta^2}$ (all other $\{\sigma_i^{-2}\}_{i \neq i_1}$ are set to infinity)

$\mathscr{B} = \{i_1\}$

Compute $\mathbf{P}$ (Eq. (4.8)), $\vec{\mu}$ (Eq. (4.7)) (initially scalars) and estimate A from Eq.(4.5)
repeat
  for $i = 1, \ldots, N$ **do**
    Compute $\xi_i = q_i^2 - s_i$
    if $\xi_i > 0$ and $\sigma_i^{-2} < \infty$ **then**
      re-estimate σ_i^{-2}
    else if $\xi_i > 0$ and $\sigma_i^{-2} = \infty$ **then**
      add i-th basis in the model ($\mathscr{B} \leftarrow \mathscr{B} \cup \{i\}$) and update σ_i^{-2}
    else if $\xi_i \leq 0$ and $\sigma_i^{-2} < \infty$ **then**
      delete i-th basis from the model ($\mathscr{B} \leftarrow \mathscr{B} \setminus \{i\}$) and set $\sigma_i^{-2} = \infty$
    end if
    Update $\mathbf{P}$, $\vec{\mu}$ and A (in this order)
    Update $\sigma_\eta^2 = \frac{\|\vec{g} - \mathbf{\Phi}\vec{\mu}\|^2}{N - card(\mathscr{B}) + \sum_{n \in \mathscr{B}} A^{-1} \sigma_n^{-2} \mathbf{P}_{nn}}$ {*card* denotes the cardinality of a set}
    Update $\mathbf{D}$ by performing the scaling $A\,\sigma_i^2$
  end for
until convergence

The *differential entropy* is a commonly used probabilistic measure of the uncertainty in the knowledge of the density of $\vec{x}$,

$$h(\vec{x}) = -\int p(\vec{x}) \log(p(\vec{x})) d\vec{x} = \frac{1}{2} \log\left(|\mathbf{\Psi P \Psi}^T|\right) + c = \frac{1}{2} \log(|\mathbf{P}|) + c \;, \tag{4.19}$$

where c is a constant that does not depend on the measurement matrix $\mathbf{\Phi}$. The next optimal projection $\vec{\phi}_{M+1}$ is selected in terms of minimizing the differential entropy. If the measurement matrix $\mathbf{\Phi}$ is augmented by adding $\vec{\phi}_{M+1}^T$ as its $(M+1)$-th row and by denoting with $h_{M+1}(\vec{x})$ the new differential entropy we have the following relation:

$$h_{M+1}(\vec{x}) = h(\vec{x}) - \frac{1}{2} \log\left(1 + \sigma_\eta^{-2} \vec{\phi}_{M+1}^T \mathbf{P} \vec{\phi}_{M+1}\right) \;, \tag{4.20}$$

where the estimation of σ_η^{-2} and $\mathbf{P}$ is based on the previous M measurements. From the above relation it is clear that the minimization of $h_{M+1}(\vec{x})$ is equivalent to maximizing $\vec{\phi}_{M+1}^T \mathbf{P} \vec{\phi}_{M+1}$. This can be carried out by performing an eigen-decomposition of the covariance matrix $\mathbf{P}$ and select $\vec{\phi}_{M+1}$ to be the eigenvector corresponding to the largest eigenvalue.

In the following, we evaluate the performance of the proposed GSM-based CS algorithm in terms of the resulting reconstruction error and achieved sparsity, by applying it on synthetic signals as well as on real-world images. In addition, we compare with the performance of other state-of-the-art norm-based and Bayesian CS methods.

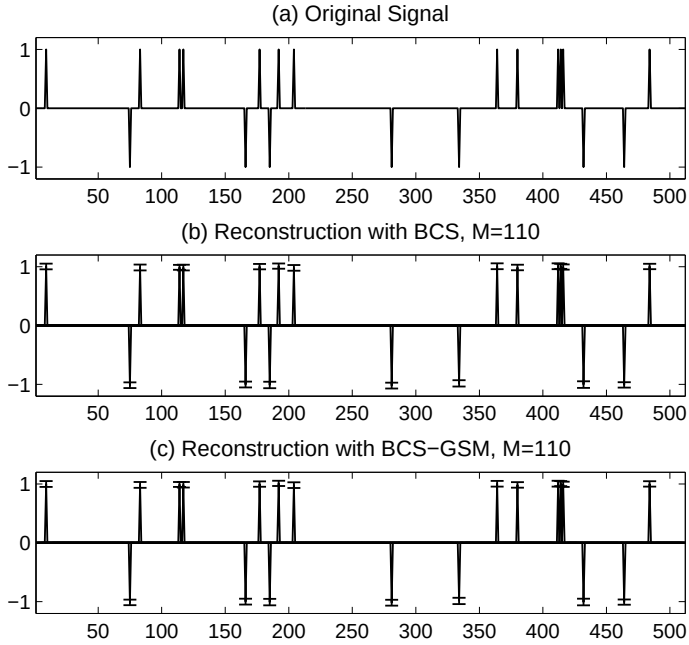


Figure 4.1: Reconstruction of uniform spikes for $N = 512$, $L = 20$, $M = 110$. (a) Original signal, (b) Reconstruction with BCS, $\|\vec{x} - \vec{x}_{BCS}\|_2/\|\vec{x}\|_2 = 0.0114$, (c) Reconstruction with BCS-GSM, $\|\vec{x} - \vec{x}_{BCS-GSM}\|_2/\|\vec{x}\|_2 = 0.0106$.

4.5 Performance evaluation: 1-D synthetic signals

In this section, we evaluate the performance of the proposed BCS-GSM method by employing synthetic signals and compare it with the performance of the standard BCS method[3], as well as with recently introduced state-of-the-art CS techniques. In the following, the reconstruction performance of a CS method is evaluated using the relative reconstruction error (RRE) defined as follows:

$$\text{relative reconstruction error} = \frac{\|\vec{x} - \vec{x}_{\text{method}}\|_2}{\|\vec{x}\|_2} \ , \tag{4.21}$$

where $\vec{x}_{\text{method}}$ is the reconstructed sparse vector obtained using a given CS method.

As a first illustration, we consider a set of simulated signals of length $N = 512$ that contain $L = 20$ spikes created by setting ± 1 at 20 locations chosen at random. The $M \times N$ measurement matrix $\mathbf{\Phi}$ is constructed by first drawing i.i.d. samples from a standard Gaussian distribution and then normalizing its columns to unit magnitude. The measurement noise is generated by drawing samples from a zero-mean Gaussian density with standard deviation $\sigma_\eta = 0.005$.

Figure 4.1(a) shows the original sparse signal, while Figures 4.1(b)-(c) present the reconstructed signals using BCS and BCS-GSM, respectively, using $M = 110$ measurements. As we can see, both algorithms recover accurately the 20 spikes, with the proposed BCS-GSM method achieving a slightly smaller reconstruction error compared with the error of the BCS approach. Besides, the error bars in the last two figures show that both methods reconstruct the original signal within a similar confidence interval.

The signal that we considered above has uniform spikes and it corresponds to the case of

[3]For the implementation of BCS we used the code included in the SparseLab package that is available online at `http://sparselab.stanford.edu/`.

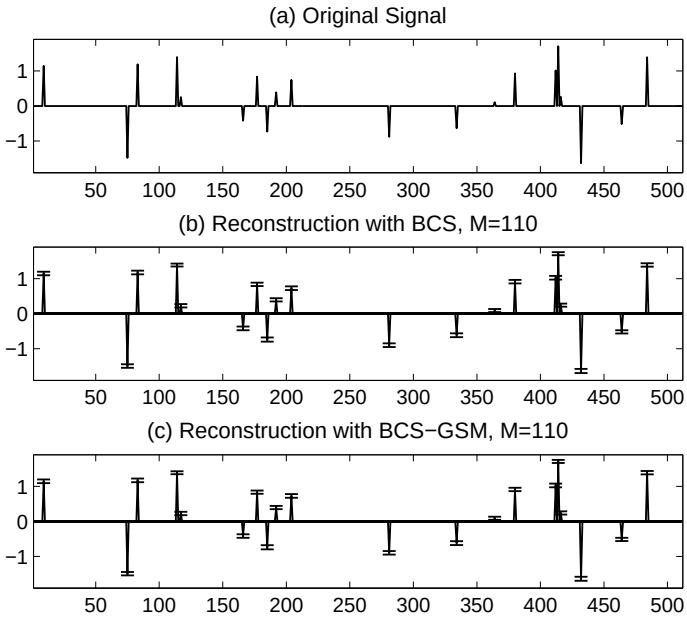


Figure 4.2: Reconstruction of non-uniform spikes for $N = 512$, $L = 20$, $M = 110$. (a) Original signal, (b) Reconstruction with BCS, $\|\vec{x} - \vec{x}_{BCS}\|_2/\|\vec{x}\|_2 = 0.0122$, (c) Reconstruction with BCS-GSM, $\|\vec{x} - \vec{x}_{BCS-GSM}\|_2/\|\vec{x}\|_2 = 0.0121$.

a sparse vector $\vec{w}$ whose non-zero components have the same magnitude. In the following, we repeat the experiment on a signal with non-uniform spikes, as shown in Figure 4.2(a). To make the comparison meaningful, the Signal-to-Noise Ratio (SNR) of both signals is fixed to be the same. The reconstructed signals using BCS and BCS-GSM, respectively, are shown in Figures 4.2(b)-(c) for $M = 110$ measurements. As in the uniform case, both algorithms recover accurately the 20 non-uniform spikes, with the proposed BCS-GSM method achieving a slightly smaller reconstruction error compared with the error of the BCS approach. Besides, the error bars in the last two figures show that the original signal is reconstructed by both methods within a similar confidence interval.

In order to test the effect of the number of CS measurements M on the performance of BCS-GSM we perform a set of 100 Monte-Carlo runs, where in each run we generate a signal with 20 randomly placed non-uniform spikes, with the value of each spike drawn from a standard Gaussian distribution. Figure 4.3 on the next page compares the reconstruction errors averaged over the 100 runs for the BCS-GSM and BCS methods, as well as the average number of non-zero components of $\vec{w}$, as a function of the number of CS measurements $M \in [60, 100]$. It is clear that the proposed scheme maintains the reconstruction performance achieved by the BCS method. On the other hand, and most importantly, there is a significant increase in the sparsity achieved by the BCS-GSM algorithm (a smaller number of non-zero components), when compared with BCS, which is equivalent to using a decreased number of basis functions (columns of $\mathbf{\Phi}$). This increase of sparsity is more apparent for a small number of CS measurements, while both methods tend to produce equally sparse vectors $\vec{w}$ as the number of CS measurements increases. This is a natural behavior, since as the number of measurements increases, the number of basis functions required to extract the information content of the signal $\vec{x}$ decreases.

Besides, the increased sparsity achieved by BCS-GSM can be exploited in reducing the storage requirements of a sensing system. In particular, instead of storing the whole signal

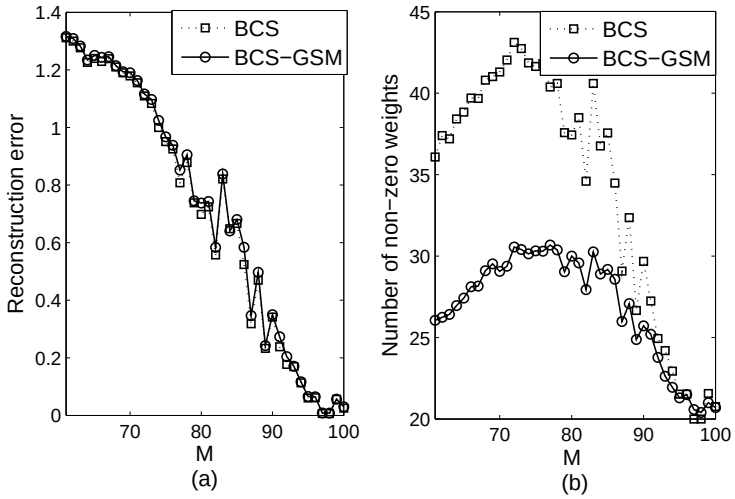


Figure 4.3: Reconstruction performance of BCS-GSM and BCS: (a) Average reconstruction error, and (b) Average number of non-zero weights, over 100 Monte-Carlo runs using random signals with 20 non-uniform spikes ($\sigma_\eta = 0.005$).

of length N, it suffices to store only the set of K indices ($K \ll N$) corresponding to the columns of $\mathbf{\Phi}$ selected by BCS-GSM. Then, assuming that $\mathbf{\Phi}$ is known, the original signal can be reconstructed from the set of K indices using the BCS-GSM method.

In order to make more meaningful the comparisons between the sparsities achieved by several CS methods, we define the *CS ratio* as follows:

$$\text{CS ratio} = \frac{\text{number of CS measurements } M}{\text{number of non-zero weights}} \ . \tag{4.22}$$

The higher the CS ratio of a CS reconstruction algorithm, the higher the achieved sparsity is for a fixed value of M.

Figure 4.4 shows the CS ratio for BCS-GSM and BCS. As it can be seen, for the same number of CS measurements M, BCS-GSM results in a higher CS ratio than the CS ratio of the BCS approach, which means that our proposed method results in a sparser solution. Besides, for both methods this ratio increases almost monotonically tending to 5 as M approaches 100, which is exactly the value of $\frac{100}{20}$, where $L = 20$ is the sparsity of the simulated signals. Figure 4.5 shows the average reconstruction error as a function of the CS ratio for the BCS-GSM and BCS methods. We observe that for the same reconstruction error, the BCS-GSM method achieves a higher CS ratio, which is equivalent to achieving a higher sparsity.

The above results provide a first indication of the effectiveness of the proposed BCS-GSM method in reconstructing sparse signals. In order to strengthen the validity of this observation we compare its performance with some of the recently introduced state-of-the-art CS methods being either deterministic (norm-based) or purely Bayesian. In particular, in the subsequent evaluations we compare with the performance of the following CS methods: 1) Gradient Projection for Sparse Reconstruction (GPSR) [34], 2) Basis Pursuit (BP) [37], 3) Stagewise Orthogonal Matching Pursuit (StOMP) [39], 4) ℓ_1-norm minimization using the primal-dual interior point method (L1EQ-PD) [147], 5) Smoothed ℓ_0 (SL0) [150], 6) BCS and 7) BCS by variational Bayes (BCS-vB) [149].[4]

[4]For the implementation of the other CS methods we used the MATLAB codes included in the packages: `http://sparselab.stanford.edu/`, `http://www.acm.caltech.edu/l1magic`, `http://www.lx.it.pt/~mtf/GPSR`, `http://ee.sharif.ir/~SLzero`, `http://www.see.ed.ac.uk/~tblumens/sparsify/sparsify.html`,

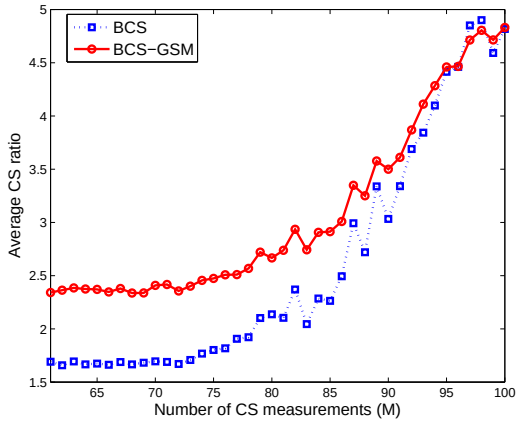


Figure 4.4: Average CS ratio as a function of M for the BCS-GSM, BCS methods ($L = 20$, $\sigma_\eta = 0.005$).

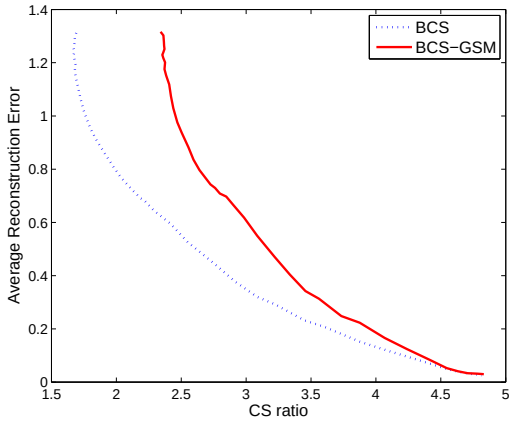


Figure 4.5: Average reconstruction error as a function of CS ratio for the BCS-GSM, BCS methods ($L = 20$, $\sigma_\eta = 0.005$).

More specifically, we are interested in testing how the reconstruction efficiency of the above CS methods is affected by the degree of sparsity, as well as by the input SNR. For this purpose, first we present the reconstruction performance of the selected CS methods on a set of random signals with uniform spikes (± 1) generated by setting $N = 512$, $L \in \{10, 15, 20\}$, SNR $\in \{10 : 5 : 35\}$ (dB) and $M = 110$. Also note that the results are averaged over 100 Monte-Carlo runs. Besides, the noise variance σ_η^2 is determined by the following relation as a function of SNR, L and N:

$$\text{SNR} = 10 \cdot \log_{10}\left(\frac{\frac{L}{N}}{\sigma_\eta^2}\right) , \tag{4.23}$$

where $\frac{L}{N}$ is the empirical estimation of the variance of the sparse vector.

Figures 4.6(a), 4.6(c), 4.6(e) on page 54 show the average RREs, for the three degrees of sparsity, as a function of SNR for the BCS-GSM and the other selected CS methods. We can see that as L decreases (i.e., sparsity increases) the Bayesian CS methods result in an increased reconstruction performance for increasing SNR values, when compared with the norm-based methods. On the other hand, and most importantly, the proposed BCS-GSM method results in enormous compression rates in terms of yielding much higher CS ratios, as shown in Figures 4.6(b), 4.6(d), 4.6(f). In particular, BCS-GSM achieves an increase of the CS ratio (or equivalently a decrease of the number of significant basis functions) even at the order of ~ 50 times in comparison with the majority of norm-based methods (cf. Fig. 4.6(f)), which is more prominent for smaller values of the SNR. Besides, notice that in contrast to the other CS methods, the CS ratios of BCS-GSM and BCS converge exactly to the corresponding true fraction CS ratio $= \frac{M}{L} = \frac{110}{L}$ for the three values of L.

In addition, we observe that the CS ratios of BCS-GSM and StOMP present a relatively flat behavior for all the SNR values. This happens because these two methods perform a correction of the noise variance and thus they refine their adaptation to the true SNR value (note the term σ_η^2/A in the definition of the matrix $\mathbf{C}$ in Eq. (4.10)). On the other hand, BCS and GPSR start with an under- or over-estimated noise variance, which is preserved until their convergence. Thus, for low SNR values a potential error in the initial estimate of σ_η^2 is important, while it does not play a significant role for large SNR values ($\sigma_\eta^2 \ll 1$) and thus resulting in an increased sparsity (high CS ratio).

`http://people.ee.duke.edu/~lihan/cs/`.

Besides, we can see that the CS ratios corresponding to BP, L1EQ-PD, SL0 and BCS-vB take a constant value $\simeq 0.21$, which is exactly the ratio $\frac{M}{N} = \frac{110}{512}$. This means that these four methods employ all the basis vectors to estimate the sparse vector. The reason for this characteristic is ought to the implementation of these methods. In contrast to BCS-GSM, StOMP, BCS and GPSR, which start from an all-zero vector and activate (or de-activate) components in each iteration, the last four methods perform the CS reconstruction by solving linear systems of equations employing the whole measurement matrix, followed by a kind of thresholding that suppresses the non-significant components. However, due to numerical inaccuracies this process usually does not result in exactly zero components, but in components with negligible amplitude. For this purpose, the computation of the CS ratio considers even the locations of negligible amplitude to be activated, yielding a small CS ratio value.

A second series of experiments is performed on a set of random signals with random non-uniform spikes drawn from a standard Gaussian distribution and normalized such that the variance of the sparse signal equals to $\frac{L}{N}$. As before, the signals are generated by setting $N = 512$, $L \in \{10, 15, 20\}$, SNR $\in \{10 : 5 : 35\}$ (dB) and $M = 110$, while the noise variance σ_η^2 is determined by Eq. (4.23). As in the uniform case, the results are averaged over 100 Monte-Carlo runs.

Figures 4.7(a), 4.7(c), 4.7(e) on page 55 show the average RREs, for the three degrees of sparsity, as a function of SNR for the BCS-GSM and the other selected CS methods. Similarly to the uniform case, we observe that as L decreases (i.e., sparsity increases) the Bayesian CS methods result in an increased reconstruction performance for increasing SNR values, when compared with the norm-based methods. However, for $L = 20$ the reconstruction is more accurate in the non-uniform case. On the other hand, the proposed BCS-GSM method results again in significantly increased compression rates in terms of yielding much higher CS ratios, as shown in Figures 4.7(b), 4.7(d), 4.7(f). In particular, BCS-GSM achieves an increase of the CS ratio (or equivalently a decrease of the number of significant basis functions) even at the order of ~ 55 times in comparison with the majority of norm-based methods (cf. Fig. 4.7(f)), which is more prominent for smaller values of the SNR. Besides, notice that in contrast to the other CS methods, the CS ratios of BCS-GSM and BCS applied on non-uniform sparse signals converge more closely to the corresponding true fraction CS ratio $= \frac{M}{L} = \frac{110}{L}$ for the three values of L.

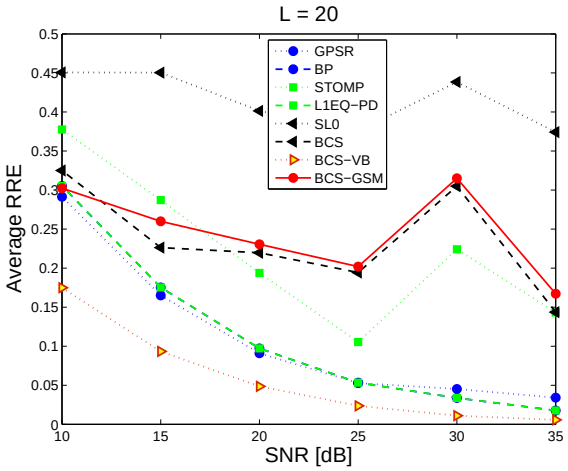
In addition, similar observations can be extracted for the specific behavior of each method (e.g., the relatively flat behavior of BCS-GSM and StOMP, or the small CS ratio values for BCS-vB, SL0 and BP) as in the case of uniform spikes.

As discussed in Section 4.4, the Bayesian framework permits the adaptive design of the measurement matrix $\mathbf{\Phi}$ with the goal of reducing the uncertainty in reconstructing the sparse vector. We compare the performance of the standard BCS-GSM implementation with its adaptive version by employing a set of 100 Monte-Carlo runs, where in each run we generate a signal of length $N = 512$ with 20 randomly placed non-uniform spikes drawn from a standard Gaussian distribution. The measurement vector $\vec{g}$ is corrupted by additive zero-mean Gaussian noise with standard deviation $\sigma_\eta = 0.005$. The initial 60 measurements are gathered by using the associated random projections, while the remaining 40 measurements are gathered sequentially based on the adaptive optimal selection of the next projection. After each new projection vector $\vec{\phi}_{M+1}$ is determined the associated reconstruction error is also computed as the average over the 100 Monte-Carlo runs.

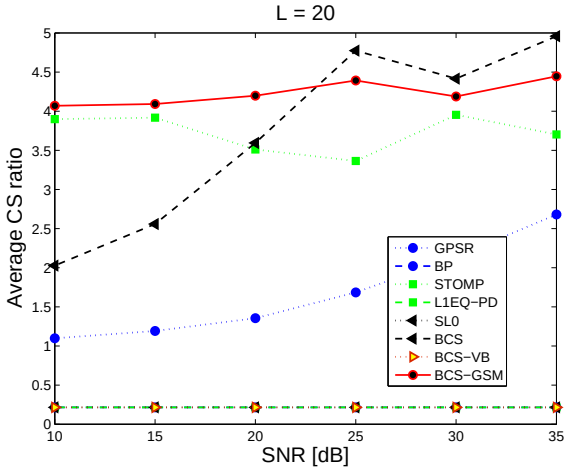
Figure 4.8 on page 56 shows the average reconstruction error of the standard (non-adaptive) and the adaptive BCS-GSM method. As it can be seen, the adaptive implementation outperforms the non-adaptive one, and particularly, the improvement achieved by the adaptive BCS-GSM increases as the number of CS measurements increases. This means that the sequential selection of the next optimal projection results in a decreased reconstruction error and this increased performance accumulates over the number of measurements, resulting in an overall

superior performance when compared to the standard BCS-GSM scheme.

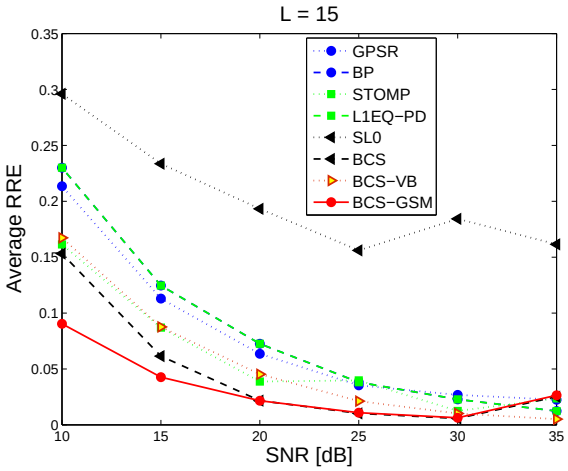
As a second illustration, we compare the reconstruction performance of the adaptive versions of BCS and BCS-GSM on a set of random sparse vectors of length $N = 512$ with $L = 15$ uniform spikes corrupted by additive Gaussian noise with $\sigma_\eta = 0.01$. Figure 4.9(a) on page 56 shows the average RREs for the two methods over 100 Monte-Carlo runs. In each run the initial 80 measurements are gathered by using the associated random projections, while the remaining 40 measurements are gathered sequentially based on the adaptive optimal selection of the next projection. The proposed BCS-GSM outperforms the standard BCS method as the number of measurements M decreases, whereas both methods achieve the same performance as M approaches 120. In addition, for fewer CS measurements BCS-GSM results in a higher CS ratio, too, as shown in Figure 4.9(b), while as M increases both methods converge to the true ratio $8 = \frac{M=120}{L=15}$.



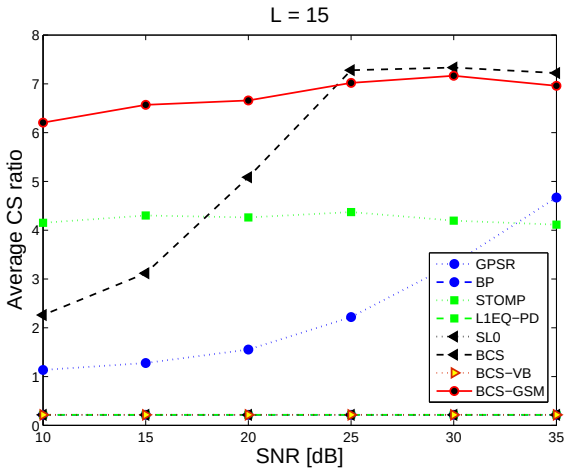
(a) RREs, uniform spikes: ($N = 512$, $L = 20$, $M = 110$)



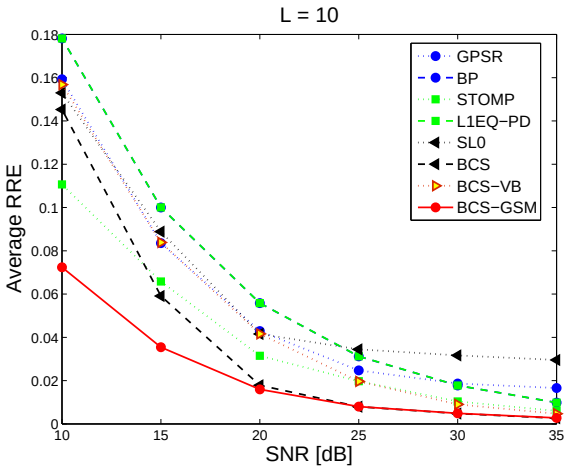
(b) CS ratios, uniform spikes: ($N = 512$, $L = 20$, $M = 110$)



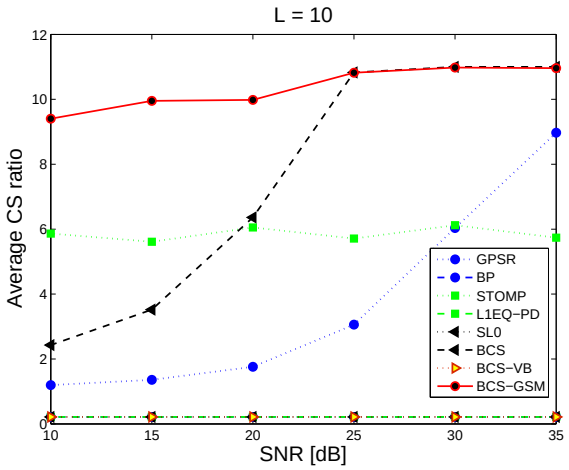
(c) RREs, uniform spikes: ($N = 512$, $L = 15$, $M = 110$)



(d) CS ratios, uniform spikes: ($N = 512$, $L = 15$, $M = 110$)

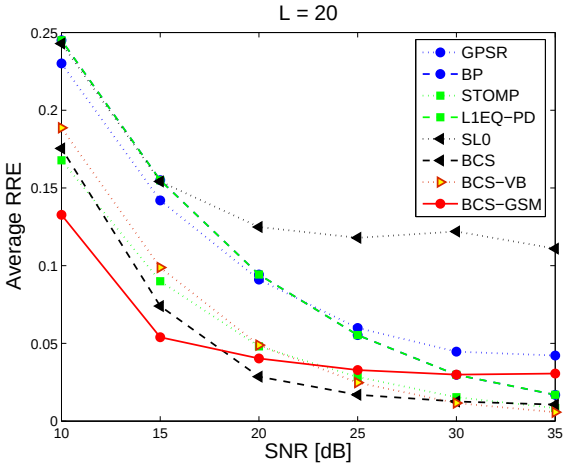


(e) RREs, uniform spikes: ($N = 512$, $L = 10$, $M = 110$)

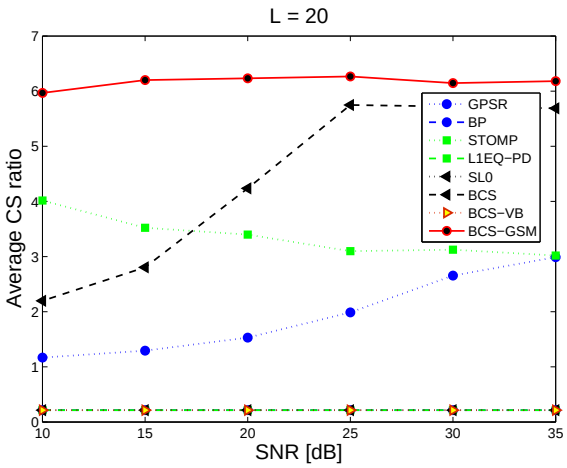


(f) CS ratios, uniform spikes: ($N = 512$, $L = 10$, $M = 110$)

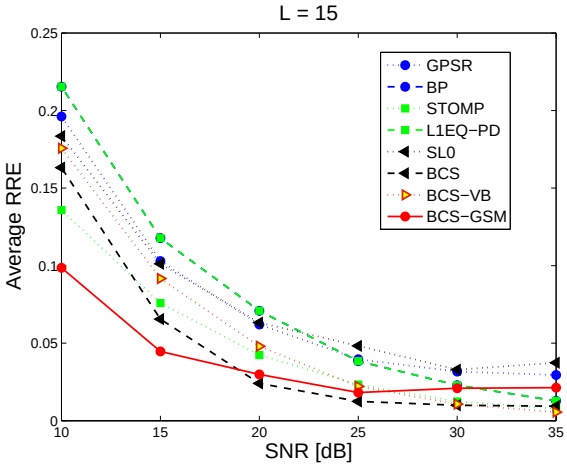
Figure 4.6: Average RREs and CS ratios as a function of SNR for the BCS-GSM and the selected CS methods for sparse signals with uniform spikes.



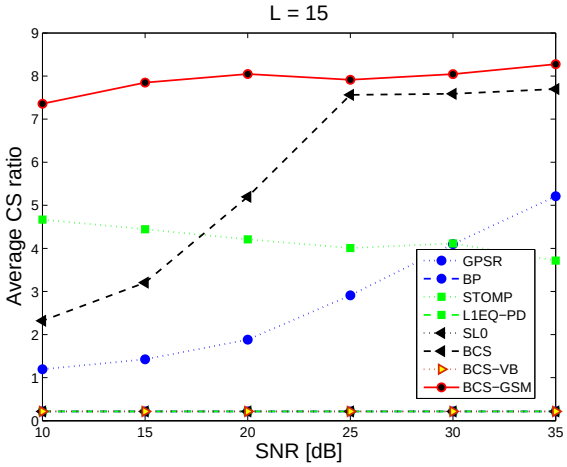
(a) RREs, non-uniform spikes: ($N = 512$, $L = 20$, $M = 110$)



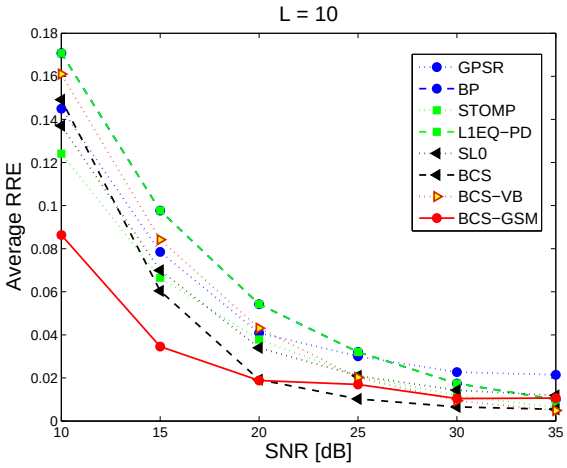
(b) CS ratios, non-uniform spikes: ($N = 512$, $L = 20$, $M = 110$)



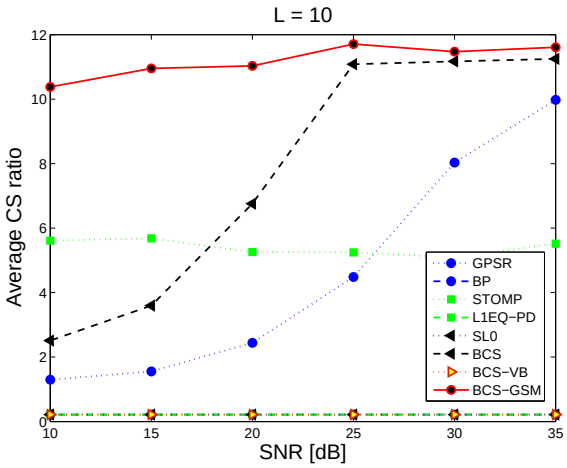
(c) RREs, non-uniform spikes: ($N = 512$, $L = 15$, $M = 110$)



(d) CS ratios, non-uniform spikes: ($N = 512$, $L = 15$, $M = 110$)



(e) RREs, non-uniform spikes: ($N = 512$, $L = 10$, $M = 110$)



(f) CS ratios, non-uniform spikes: ($N = 512$, $L = 10$, $M = 110$)

Figure 4.7: Average RREs and CS ratios as a function of SNR for the BCS-GSM and the selected CS methods for sparse signals with non-uniform spikes.

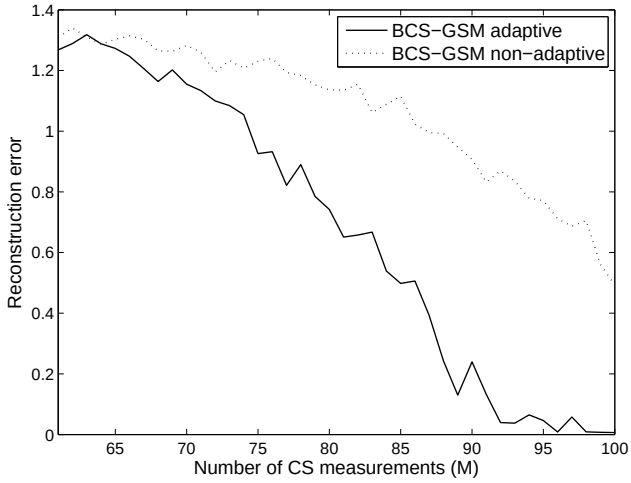
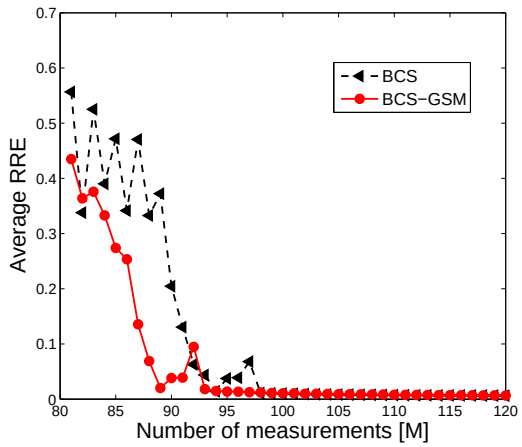
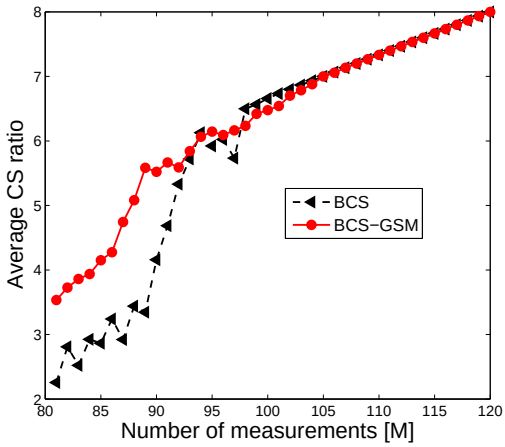


Figure 4.8: Average reconstruction error of the standard (non-adaptive) and adaptive BCS-GSM, over 100 Monte-Carlo runs using random signals with 20 spikes ($\sigma_\eta = 0.005$).



(a) Average reconstruction error of the adaptive BCS-GSM and BCS methods over 100 Monte-Carlo runs using random signals with $L = 15$ uniform spikes ($N = 512$, $\sigma_\eta = 0.01$).



(b) Average CS ratios of the adaptive BCS-GSM and BCS methods over 100 Monte-Carlo runs using random signals with $L = 15$ uniform spikes ($N = 512$, $\sigma_\eta = 0.01$).

Figure 4.9: Average RREs and CS ratios over 100 Monte-Carlo runs for the adaptive BCS-GSM and BCS methods using random signals with $L = 15$ uniform spikes ($N = 512$, $\sigma_\eta = 0.01$).

4.6 Performance evaluation: Image reconstruction

In the second set of experiments, the performance of our proposed BCS-GSM algorithm is evaluated and compared with the performance of other CS methods for the reconstruction of images with distinct content. As a first illustration, we examine the validity of the proposed BCS-GSM method in reconstructing images by comparing its performance with the standard BCS method [55], as well as with the norm-based optimization schemes BP [37] and StOMP [39] (combined with a CFAR thresholding scheme), which achieved the best reconstruction performance on the three 256×256 (noise-free) images of distinct content shown in Figure 4.10, along with their noisy versions obtained by adding zero-mean Gaussian noise with SNR $= 5, 10$ dB. Each image is sparsified by a transformation in the 2-D Discrete Wavelet Transform (DWT) domain (cf. Chapter 2). In particular, the images are decomposed in 6 scales using the Daubechies' wavelet with 4 vanishing moments ("db4"). The CS measurements $\vec{g}$ are acquired by applying Hadamard matrices $\mathbf{\Phi}$ on the wavelet coefficients vector $\vec{w}$. In addition, the quality of the reconstructed image (of size $P \times Q$) is measured via the Peak Signal-to-Noise Ratio (PSNR), which is defined as follows (in dB):

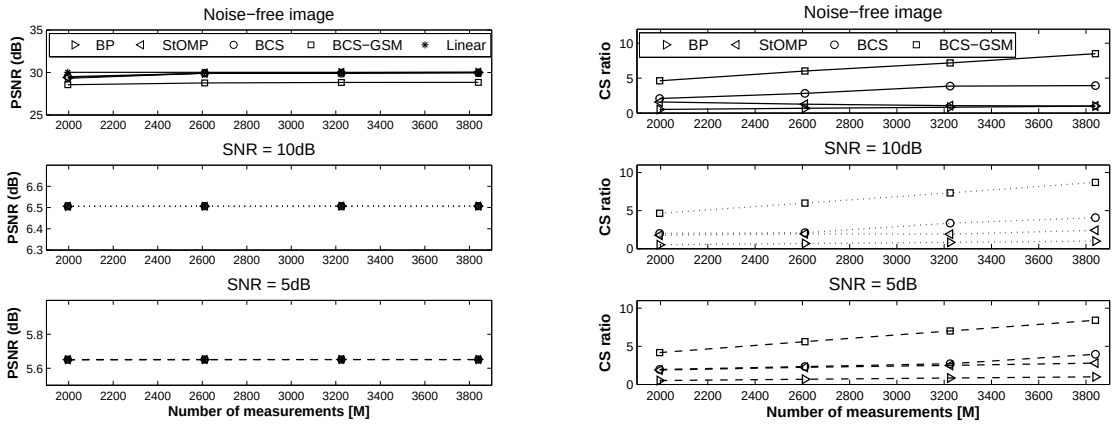
$$\text{PSNR} = 20 \cdot \log_{10} \left(\frac{\max\{I\}}{\sqrt{\frac{1}{PQ} \sum_{p=1}^{P} \sum_{q=1}^{Q} |I(p,q) - \hat{I}(p,q)|^2}} \right) , \tag{4.24}$$

where I and $\hat{I}$ denote the original and reconstructed image, respectively, $\max\{I\}$ is the maximum pixel value of image I and $I(p,q)$ is the pixel value at the position (p,q).

Figure 4.10: Test images.

Figure 4.11(a) shows the PSNRs between the reconstructed noise-free/noisy images and the original (noise-free) "*Indor 2*" image, using BCS-GSM, linear reconstruction (inverse DWT), which achieves the best performance, and the selected norm-based CS methods for the two SNR values as a function of the number of measurements, where the number of CS measurements is equal to the sum of a portion c of the detail coefficients of the decomposition levels $3-6$ plus the coarse-scale approximation coefficients. The value of c varies in $\{0.4{:}0.2{:}1\}$. As it can be seen, the proposed BCS-GSM method achieves similar reconstruction performance with the other four methods for the same number of measurements, with a decrease of PSNR which is at most $\sim$1 dB in the noise-free case, while there is no difference in the two noisy cases.

Most importantly, Figure 4.11(b) shows the CS ratio for each one of the four CS methods. We can see that the proposed BCS-GSM method results in a much sparser representation of the original image as M increases, by reducing the number of significant basis functions by as much as 50%. Thus, Figures 4.11(a)-4.11(b) indicate that the proposed scheme could reduce significantly the storage requirements of a large image database, while maintaining a high reconstruction quality. This can be very significant in applications such as image classification and retrieval.



(a) PSNRs for "*Indor 2*" image as a function of the number of CS measurements for SNR = 5, 10 dB.

(b) CS ratios of the selected CS recovery methods for "*Indor 2*" image as a function of the number of CS measurements for SNR = 5, 10 dB.

Figure 4.11: PSNRs and CS ratios for "*Indor 2*" image as a function of the number of CS measurements for SNR = 5, 10 dB.

On the other hand, Figure 4.12 shows the PSNR and CS ratio values for the BCS-GSM and BCS methods along with their adaptive versions applied on "*Indor 2*" image by varying the number of CS measurements M in the interval $[2513, 2611]$. The adaptive implementation achieves the same reconstruction performance in terms of PSNR for both methods, whereas the CS ratio increases, which is equivalent to a more efficient sparse representation (3.03% for the BCS-GSM method on average).

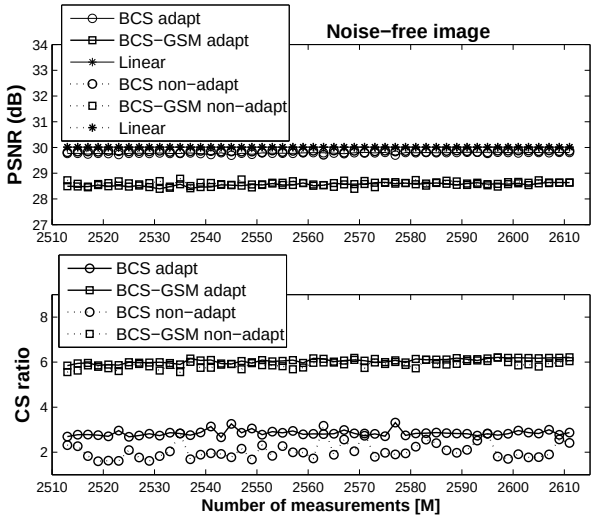


Figure 4.12: PSNR and CS ratios of the BCS-GSM, BCS and their adaptive implementations for "*Indor 2*" image.

Table 4.1 shows the PSNRs and CS ratios of the four CS methods for the other two test images by setting $c = 0.6$, which corresponds to $M = 2611$ measurements. As for the "*Indor 2*" image, the proposed BCS-GSM approach achieves a significantly increased sparsity for representing images with distinct contents (with a maximum increase of sparsity over 50%), while yielding the same reconstruction performance in the noise-free and the two noisy cases.

IMAGE		BP		StOMP		BCS		**BCS-GSM**		Linear
		PSNR	CS ratio	PSNR	CS ratio	PSNR	CS ratio	PSNR	CS ratio	PSNR
Indor 4	Noise-free	29.27	0.68	29.16	1.85	29.04	3.63	28.35	5.63	29.60
	10 dB	7.08	0.68	7.09	1.87	7.09	2.05	7.08	5.49	7.09
	5 dB	6.15	0.67	6.15	2.45	6.15	4.31	6.14	5.95	6.15
Nemasup	Noise-free	25.43	0.67	25.34	2.11	25.35	2.24	25.32	4.13	25.61
	10 dB	9.95	0.68	9.96	2.92	9.96	1.90	9.95	4.30	9.96
	5 dB	5.57	0.68	5.56	2.51	5.57	2.01	5.56	4.15	5.57

Table 4.1: Performance comparison in terms of PSNR and CS ratio values for the reconstruction of "*Indor 4*" and "*Nemasup*" images (noise-free & noisy versions) with $c = 0.6$ ($M = 2611$).

In a second series of experiments, we evaluate the performance of BCS-GSM by applying it on a set of six medical images of size 128×128, which are shown in Figure 4.13. As before, the images are sparsified in the 2-D DWT domain by decomposing them in 5 scales using the "db4" wavelet. Except for the original (noiseless) images we generate two noisy versions of them by adding zero-mean Gaussian noise resulting in SNR $= 7.5, 15$ dB. As mentioned in Chapter 2, the detail wavelet coefficients represent the high-frequency content of a given signal and they are characterized by a highly sparse behavior, whereas the approximation coefficients correspond to a coarse representation of it. Thus, the CS algorithms are applied on the detail coefficients only and the reconstruction of the original signal is performed by adding the approximation coefficients to the reconstructed signal obtained from the detail coefficients.

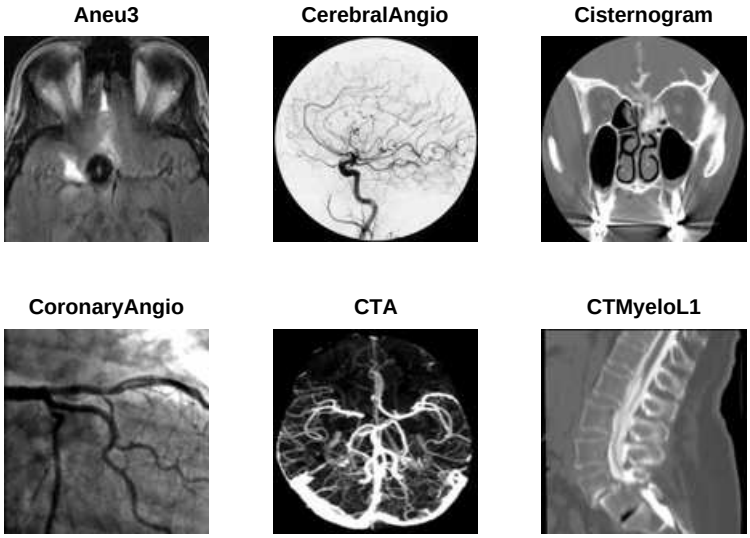


Figure 4.13: Medical images (128×128) used for evaluation of the performance of BCS-GSM.

In the subsequent experiments we apply the several CS algorithms using a portion of the detail coefficients. In particular, if N_{detail} is the number of the detail coefficients, we evaluate the performance using a subset of size $c \cdot N_{detail}$ with $c \in \{0.3, 0.4, 0.5, 0.65\}$ (or equivalently by employing 30%, 40%, 50% and 65% of the detail coefficients). The proposed BCS-GSM method is compared with the following CS techniques: 1) standard BCS, 2) BP, 3) StOMP (combined with a CFAR thresholding scheme), 4) ℓ_1-norm minimization using the primal-dual interior point method (L1EQ-PD) and 5) the linear reconstruction, which is simply the inverse

2-D DWT and gives the optimal reconstruction.

Figures 4.14, 4.15 on the facing page show the PSNRs between the reconstructed (noiseless and noisy) images and the corresponding original (noiseless) image, for the BCS-GSM, as well as for the other five reconstruction approaches, as a function of the number of measurements for the two SNR values. First, we observe that for the six images the reconstruction performance of all methods decreases as the SNR decreases, something that we expected. However, it is clear that the proposed BCS-GSM method achieves practically the same PSNR with the selected CS methods and the optimal linear reconstruction. In particular, the difference in PSNR with the linear reconstruction is less than 1 dB in the noiseless case, while it is negligible in the two noisy cases.

Besides, the increased number of measurements in the selected range $M \in [1900, 4200]$ does not affect the reconstruction PSNR as much as one would expect. The reason for this, almost flat, behavior is the high sparsity of the selected images in the wavelet domain. Since they consist of lines and edges on a relatively homogeneous background the most significant coefficients of their corresponding 2-D DWT transform are placed in the finer (high-frequency) scales. Thus, the first $M = 2000$ measurements generated by employing 30% ($c = 0.3$) of the detail coefficients already carry the most significant information content. The addition of more measurements, that is, of more detail coefficients, improves only slightly (in practice negligibly) the reconstruction performance. However, even though the PSNR remains almost constant, the CS ratio profits by the increased number of measurements, yielding higher values and thus an increased sparsity.

The increased capabilities of BCS-GSM in providing a highly sparse representation in the case of images becomes more apparent in Figures 4.16, 4.17 on page 62, which present the corresponding CS ratio values. Obviously, BCS-GSM outperforms all the other CS methods, increasing the sparsity of the representation by as much as 15 times as the number of measurements increases. In addition, this significantly increased performance is robust even in the low-SNR case. For a visual inspection of the reconstruction performance, Figures 4.18-4.21 show the original and reconstructed images corresponding to "Aneu3", "Cisternogram", "CoronaryAngio" and "CTMyeloL1" by setting $c = 0.5$ for the two SNR values.

In the following section, BCS-GSM is extended in the case where multiple observations of the original signal are available. The proposed method is developed in a distributed way and thus it is amenable to an implementation by the nodes of a sensor network.

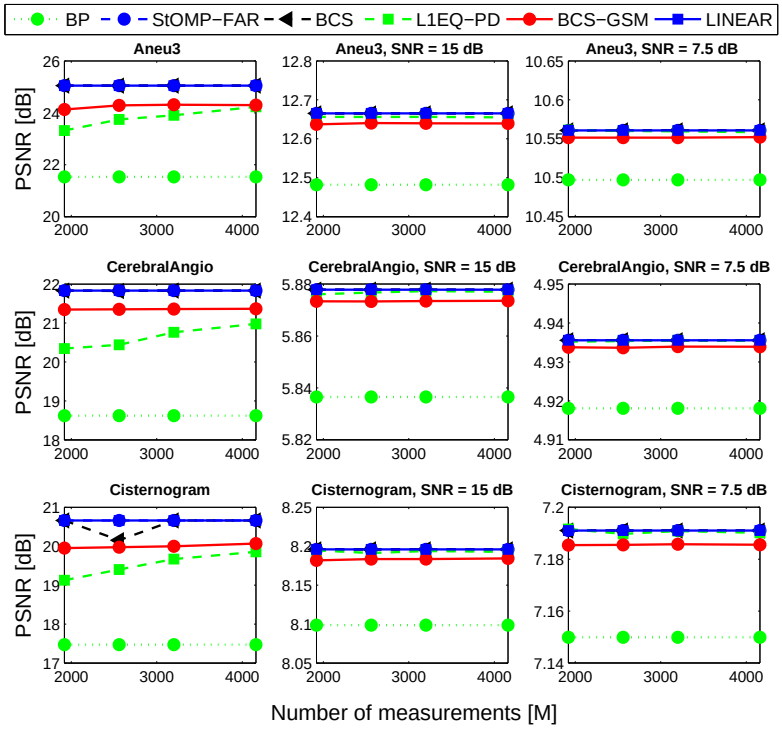


Figure 4.14: PSNRs for CS reconstruction of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of the number of measurements and for SNR $= 7.5, 15$ dB.

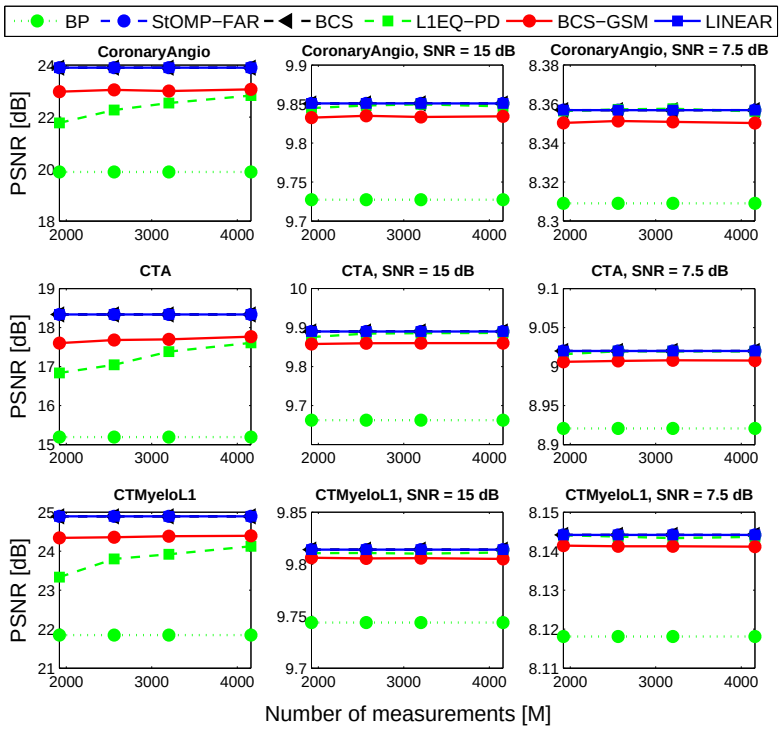


Figure 4.15: PSNRs for CS reconstruction of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of the number of measurements and for SNR $= 7.5, 15$ dB.

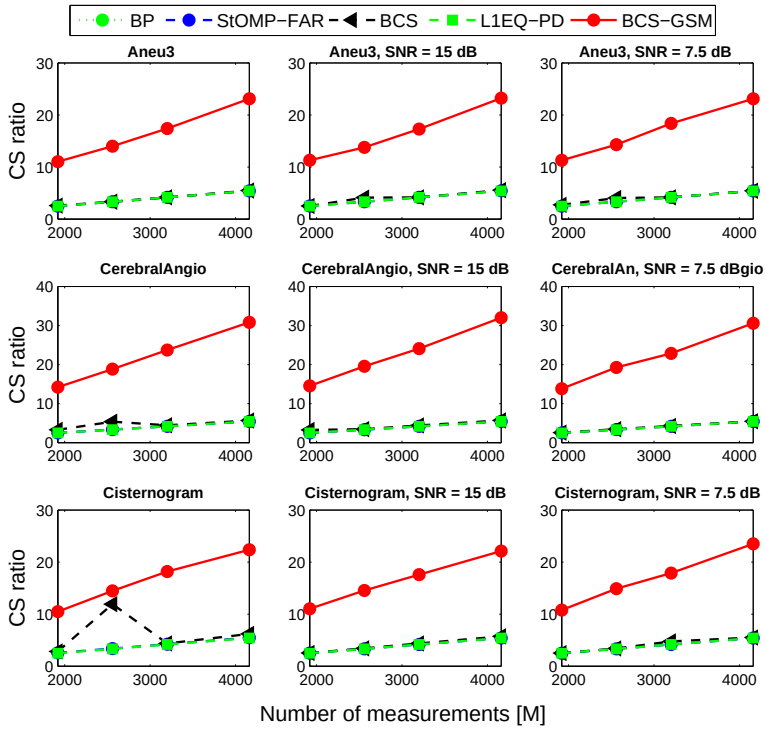


Figure 4.16: CS ratios for CS reconstruction of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of the number of measurements and for SNR $= 7.5, 15$ dB.

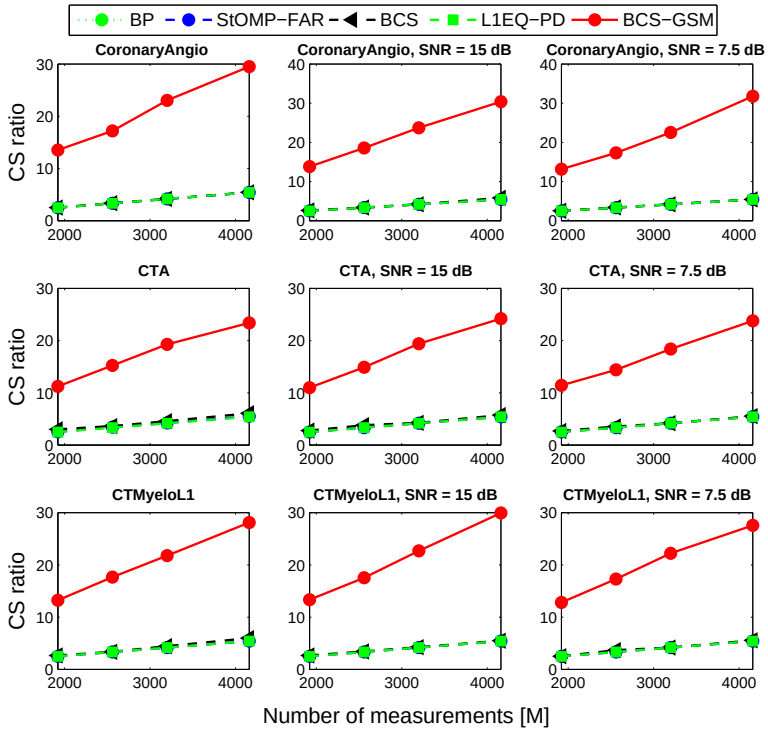


Figure 4.17: CS ratios for CS reconstruction of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of the number of measurements and for SNR $= 7.5, 15$ dB.

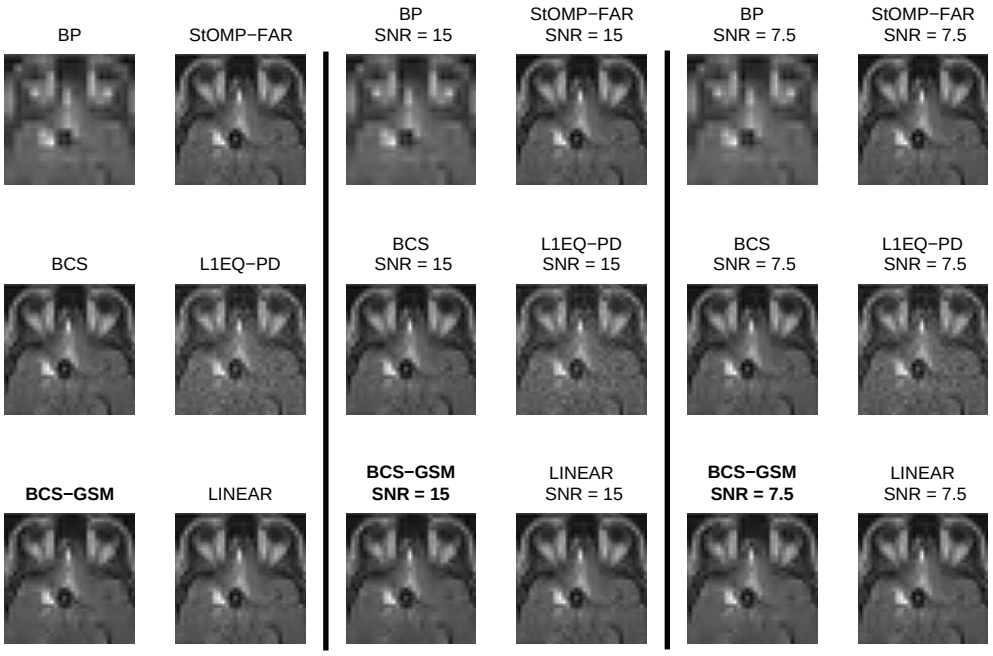


Figure 4.18: Original and CS reconstructed images of "Aneu3" for SNR $= 7.5, 15$ dB.

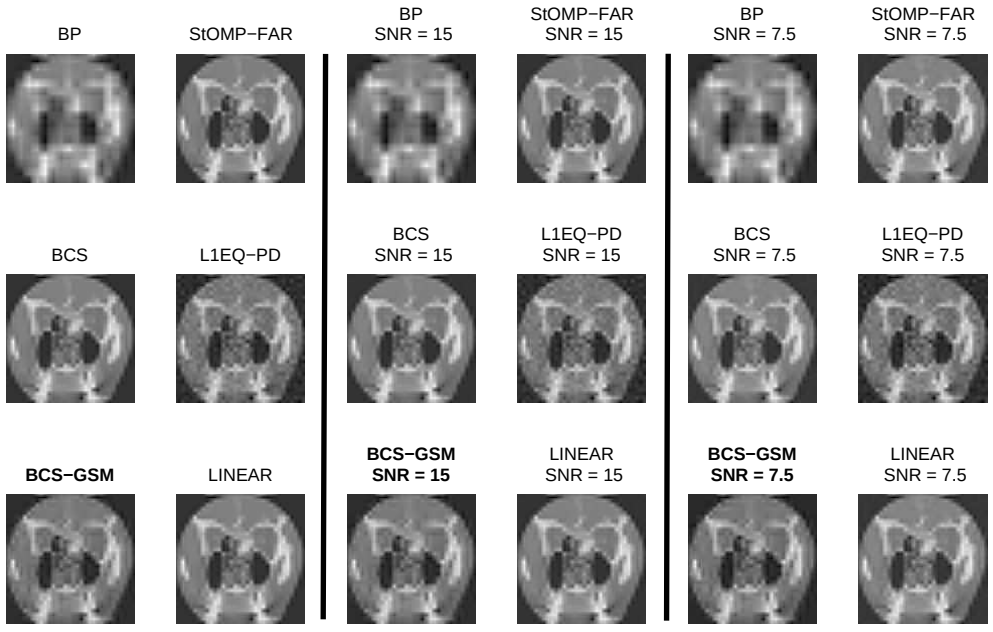


Figure 4.19: Original and CS reconstructed images of "Cisternogram" for SNR $= 7.5, 15$ dB.

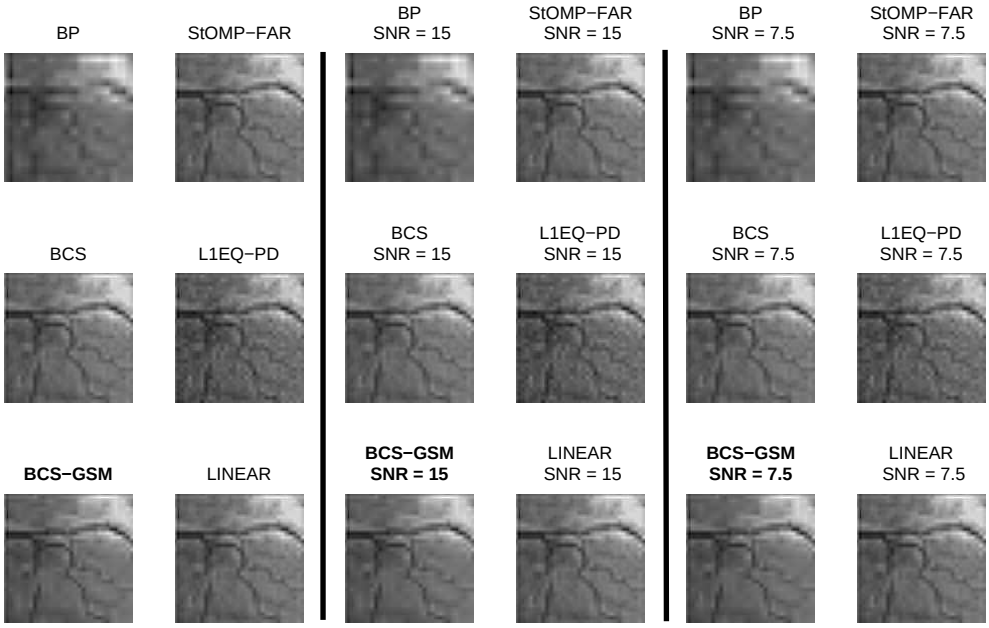


Figure 4.20: Original and CS reconstructed images of "CoronaryAngio" for SNR $= 7.5, 15$ dB.

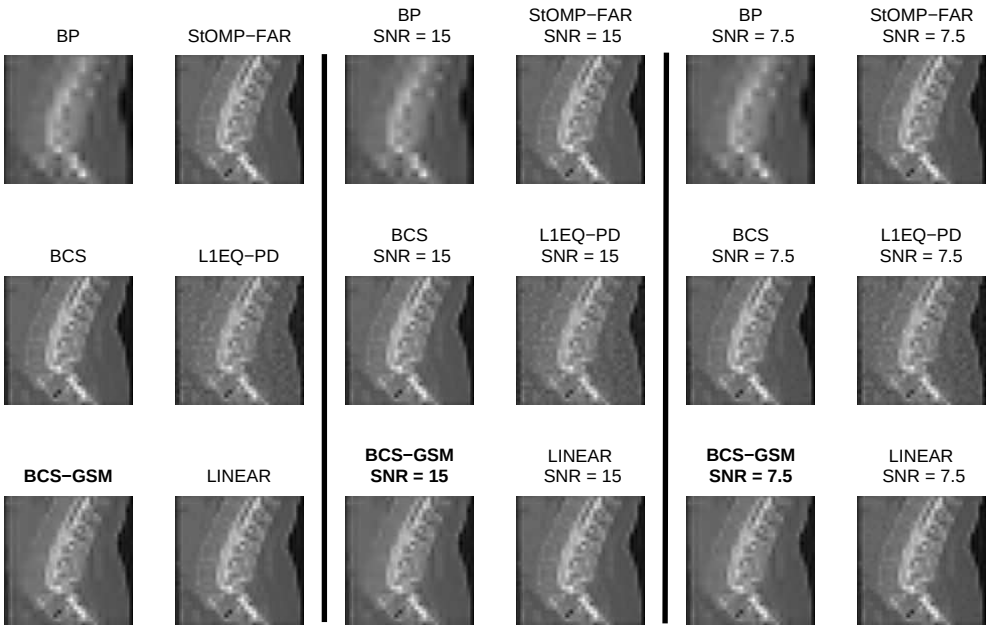


Figure 4.21: Original and CS reconstructed images of "CTMyeloL1" for SNR $= 7.5, 15$ dB.

4.7 Multiple-measurement BCS using a GSM

CS theory provides a paradigm for simultaneously sensing and compressing a signal using a small subset of random incoherent projection coefficients, enabling a potentially significant reduction in the sampling and computation costs at a sensing system with limited capabilities. The several CS methods presented so far provide sparse representations and reconstruction for the *single measurement vector (SMV)* case. In particular, a signal having a sparse representation in a transform basis $\mathbf{\Psi}$ can be reconstructed from a single vector containing a small number of projections onto a measurement basis $\mathbf{\Phi}$, which is incoherent with the first one. The property of *asymmetry* of the CS-based approaches is also a crucial point for the design of real-time sensing systems, since the compression part is of very low complexity (simple linear projections), while the main computational burden is on the decompression part where increased processing capabilities and computational resources are available.

However, there are cases where a single sensing device or a number of sensing devices acquires multiple measurements of the same original signal resulting in an *ensemble of signals*. Recent studies extended the CS framework by presenting methods for the sparse representation and reconstruction of a signal ensemble whose individual signals are characterized by a *common sparsity structure* [62, 73]. In this case, the multiple measurement vectors are obtained by projecting each signal of the ensemble on the same over-complete dictionary.

The problem of sparse representation and reconstruction in the case of *multiple measurement vectors (MMV)* in an over-complete dictionary is motivated by several inverse problems that arise in distinct fields, such as in astronomy [90], medical imaging [88] and multi-view imaging [122]. In the following sections of this chapter we extend the SMV BCS-GSM algorithm described above, for reconstructing the original signal based on a *set of noisy CS measurements vectors*, where each vector results by projecting the original signal on *distinct over-complete dictionaries*.

In the case of multiple measurement vectors the corresponding noisy CS reconstruction problem is stated as follows:

$$\mathbf{G} = \mathbf{\Phi W} + \mathbf{H} \ , \tag{4.25}$$

where $\mathbf{G} = [\vec{g}_1, \ldots, \vec{g}_K] \in \mathbb{R}^{M \times K}$ is the MMV matrix, $\mathbf{\Phi} = [\vec{\phi}_1, \ldots, \vec{\phi}_M]^T \in \mathbb{R}^{M \times N}$ is a random measurement matrix (over-complete dictionary), $\mathbf{W} = [\vec{w}_1, \ldots, \vec{w}_K] \in \mathbb{R}^{N \times K}$ is the weight vectors matrix and $\mathbf{H} = [\vec{\eta}_1, \ldots, \vec{\eta}_K] \in \mathbb{R}^{M \times K}$ is the noise matrix. Each sparse weight vector is the transform-domain equivalent of the corresponding original space-domain signal, $\vec{x}_i = \mathbf{\Psi}_i \vec{w}_i$, $i = 1, \ldots, K$, where the columns of $\mathbf{\Psi}_i \in \mathbb{R}^{N \times N}$ correspond to the transform basis functions. In general, each $\vec{x}_i$ is considered to be sparse on a different basis $\mathbf{\Psi}_i$. For $K = 1$ the problem is reduced to the standard CS reconstruction using a single measurement vector.

Several CS methods have been introduced recently that give an estimate of $\mathbf{W}$ satisfying Eq. (4.25) by solving a norm-based constrained optimization problem [63, 64, 65, 123]. On the other hand, the work presented in [124] develops a reconstruction method in a Bayesian framework by modelling the prior belief that the majority of the rows of $\mathbf{W}$ will be zero, due to the assumption for a joint sparsity structure, by employing a zero-mean Gaussian distribution on the norm of each individual row. As in the SMV case the Bayesian framework provides the critical advantage that we obtain not only a *point estimate* of the signal, as the norm-based methods do, but also a confidence interval, which can be employed to select appropriately the future measurements such that to reduce the reconstruction uncertainty.

However, there are cases where we are interested in reconstructing a *single original signal* from multiple measurements. For instance, this is a usual case in astronomy when we try to reconstruct a certain area of the sky using multiple observations of it [90], in a multi-view imaging system where we are interested in reconstructing the scene from the measurements

acquired by the cameras [122] or in a Direction-of-Arrival (DOA) estimation problem using a sensor network, where we try to reconstruct the vector of sources' positions using multiple observations of it. For the solution of such problems we generalize the work presented in the previous sections in the case of multiple measurements vectors.

4.7.1 Multiple-measurement vectors model

In this section we present the signal model considered in the development of the proposed CS method, as well as the reconstruction algorithm by employing multiple measurement vectors. In the following, we are given the original signal $\vec{x} \in \mathbb{R}^N$ with a sparse transform coefficient vector $\vec{w}$, which is observed K times via a set of random measurement matrices $\{\mathbf{\Phi}_i\}_{i=1}^K \in \mathbb{R}^{M\times N}$. These matrices are incoherent with $\mathbf{\Psi}$ resulting in a set of K measurement vectors given by:

$$\vec{g}_i = \mathbf{\Phi}_i \vec{w} + \vec{\eta}_i \ , \ i = 1, \ldots, K \tag{4.26}$$

where $\vec{\eta}_i \in \mathbb{R}^M$ is a noise vector with unknown variance σ_η^2. The set $\{\mathbf{\Phi}_i\}_{i=1}^K$ can contain, for instance, Hadamard matrices or matrices with i.i.d. Gaussian entries. Such matrices are incoherent with any fixed transform matrix $\mathbf{\Psi}$ with high probability (universality property). Also notice that this model differs from the one given by Eq. (4.25) in that *distinct* measurement matrices are used on a *single* weight vector, instead of a matrix.

Thus, given the matrix $\mathbf{G} = [\vec{g}_1, \ldots, \vec{g}_K]$ containing the multiple measurement vectors and the set of measurement matrices $\{\mathbf{\Phi}_i\}_{i=1}^K$ the reconstruction problem reduces to estimating the sparse weight vector $\vec{w}$. In [90] this problem is solved by employing a norm-based constrained optimization approach. The reconstruction can be also recast as an SMV problem and solved with one of the many norm-based CS approaches as follows:

$$\hat{\vec{w}} = \arg\min_{\vec{w}} \|\vec{w}\|_1 \quad s.t. \quad \|\mathbf{G} - \mathbf{\Phi}\vec{w}\| < \epsilon \ , \tag{4.27}$$

where $\mathbf{\Phi} = diag(\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_K)$ is a block-diagonal matrix and ϵ is the noise level.

On the other hand, when the inversion of CS measurements is treated from a Bayesian perspective, then, given the prior belief that $\vec{x}$ is L-sparse in basis $\mathbf{\Psi}$ (that is, only L of $\vec{w}$'s components have "significant" amplitude) and the set of CS measurement vectors $\mathbf{G}$, the objective is to formulate a *posterior probability distribution* for $\vec{w}$. As in the SMV case, this could improve the accuracy over the point estimate given by a norm-based approach and provide confidence intervals in the approximation of $\vec{x}$, which can be used to guide the optimal design of additional CS measurements with the goal of reducing the uncertainty in reconstructing $\vec{x}$.

For this purpose, we extend the work presented in the previous sections of this chapter by incorporating the set of multiple measurement vectors in the reconstruction of $\vec{x}$. Under the assumption of zero-mean Gaussian noise vectors with the same variance σ_η^2, we obtain the following Gaussian likelihood model for the measurement vector $\vec{g}_i$, $i = 1, \ldots, K$,

$$p(\vec{g}_i|\vec{w},\, \sigma_\eta^2) = (2\pi\sigma_\eta^2)^{-M/2} \cdot \exp\Big(-\frac{1}{2\sigma_\eta^2}\|\vec{g}_i - \mathbf{\Phi}_i\vec{w}\|\Big) \ . \tag{4.28}$$

As mentioned above, the goal is to seeking a full posterior density function for $\vec{w}$ and σ_η^2.

4.7.2 BCS-GSM inversion using multiple measurement vectors

As in the above work, the proposed extension consists in modeling directly the prior distribution of $\vec{w}$ with a heavy-tailed distribution, which promotes its sparsity. In particular, we model directly the prior distribution of $\vec{w}$ by means of a GSM. The density of $\vec{w}$ conditioned on the

variable A is a zero-mean multivariate Gaussian given by Eq. (4.4), while the ML estimate of the variable A is obtained from Eq. (4.5). The assumption of independence yields a diagonal covariance matrix $\mathbf{\Sigma} = diag(\sigma_1^2, \dots \sigma_N^2)$.

Assuming that the noise variance σ_η^2, the value of A, and the covariance matrix $\mathbf{\Sigma}$ have been estimated, given the CS measurement vector $\vec{g}_i$ and the measurement matrix $\mathbf{\Phi}_i$, the posterior density of $\vec{w}$ is given by the Bayes' rule,

$$p(\vec{w}|\vec{g}_i, A, \mathbf{\Sigma}, \sigma_\eta^2) = \frac{p(\vec{g}_i|\vec{w}, \sigma_\eta^2)p(\vec{w}|A, \mathbf{\Sigma})}{p(\vec{g}_i|A, \mathbf{\Sigma}, \sigma_\eta^2)} \ , \tag{4.29}$$

which is a multivariate Gaussian distribution whose mean $\vec{\mu}_i$ and covariance $\mathbf{P}_i$ are given by,

$$\vec{\mu}_i = \sigma_\eta^{-2}\mathbf{P}_i\mathbf{\Phi}_i^T\vec{g}_i \ , \tag{4.30}$$

$$\mathbf{P}_i = (\sigma_\eta^{-2}\mathbf{\Phi}_i^T\mathbf{\Phi}_i + \mathbf{D})^{-1} \ , \ i = 1, \dots, K \tag{4.31}$$

where $\mathbf{D} = diag((A\sigma_1^2)^{-1}, \dots, (A\sigma_N^2)^{-1})$ is the diagonal matrix containing the scaled variances of $\vec{w}$'s components. The presence of the scale parameter A in the proposed GSM-based BCS method provides an additional degree of freedom, which may result in a more accurate modeling of the true sparsity of the signal of interest, as well as the noise component, when compared with previous BCS approaches [55, 60].

The problem of estimating the sparse weight vector $\vec{w}$ reduces to estimating the unknown model parameters $A, \mathbf{\Sigma}, \sigma_\eta^2$, by combining *type-II* ML estimations from the K measurement vectors. For this purpose, we estimate the unknown parameters σ_η^2, $\{\sigma_j^2\}_{j=1}^N$ iteratively by employing a modified version of the marginal log-likelihood function given by Eq. (4.10). In particular, we have to incorporate explicitly the information provided by the matrix $\mathbf{G}$. The most convenient and straightforward way to do this is to sum up the contributions of the individual measurement vectors $\vec{g}_i$ resulting in the following log-likelihood function:

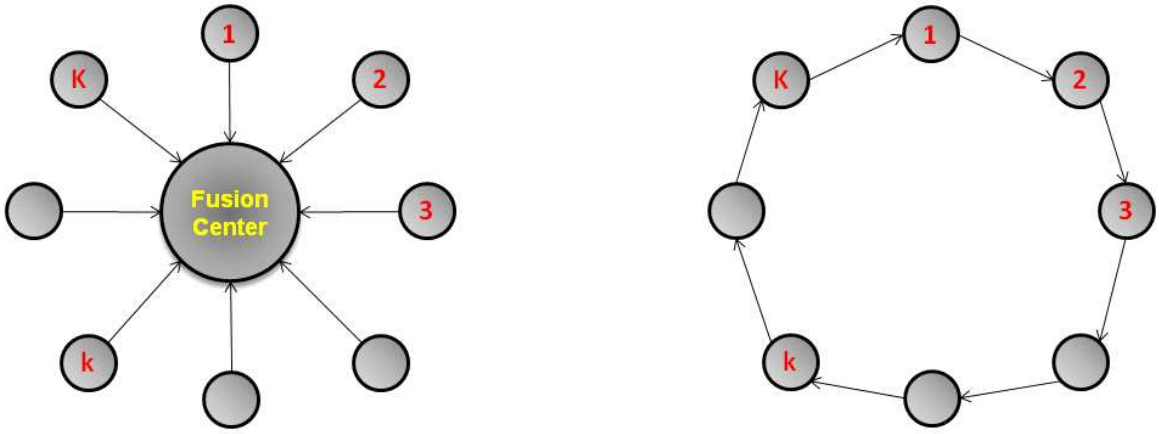
$$\begin{aligned}\mathcal{L}(\sigma_\eta^2, \{\sigma_j^{-2}\}_{j=1}^N) &= \sum_{i=1}^{K} \log[p(\vec{g}_i|A, \sigma_\eta^2, \{\sigma_j^{-2}\}_{j=1}^N)] \\ &= -\frac{1}{2}\Big[KL\ \log(2\pi) + \sum_{i=1}^{K}\log(|\mathbf{C}_i|) + \sum_{i=1}^{K}\vec{g}_i^T\mathbf{C}_i^{-1}\vec{g}_i\Big] \ ,\end{aligned} \tag{4.32}$$

where $\mathbf{C}_i = \frac{\sigma_\eta^2}{A}\mathbf{I} + \mathbf{\Phi}_i\mathbf{\Sigma}\mathbf{\Phi}_i^T$. It is clear from Eq. (4.32) that the scaling factor of A^{-1} plays an important role in the estimation process, since it controls the heavy-tailed behavior of the diagonal elements of $\mathbf{D}$ and consequently of the covariance matrices $\{\mathbf{P}_i\}_{i=1}^K$, and thus, the sparsity of the estimated weight vector $\vec{w}$ which depends on the corresponding mean vectors $\{\vec{\mu}_i\}_{i=1}^K$, as we will see in the subsequent analysis. The addition and deletion of candidate basis functions (columns of $\{\mathbf{\Phi}_i\}_{i=1}^K$) is performed with the goal to increasing monotonically the marginal likelihood. Following a similar incremental procedure as the one used in the implementation of Algorithm 1 the computational cost for updating the most "expensive" quantities of Eq. (4.32), namely the determinants $\{|\mathbf{C}_i|\}_{i=1}^K$ and the inverses $\{\mathbf{C}_i^{-1}\}_{i=1}^K$, is reduced significantly.

Besides, the additive character of Eq. (4.32) indicates that the proposed method is amenable to a distributed implementation by the nodes a sensor network, as we will discuss in the following sections.

After some algebraic manipulation we can see that the marginal likelihood is decoupled in two terms as follows:

$$\mathcal{L}(\sigma_\eta^2, \{\sigma_j^{-2}\}_{j=1}^N) = \mathcal{L}(\sigma_\eta^2, \{\sigma_j^{-2}\}_{j=1, j\neq j'}^N) + l(\sigma_{j'}^{-2}) \tag{4.33}$$



(a) Star-shaped network topology

(b) Ring-shaped network topology

Figure 4.22: Star-shaped and ring-shaped topologies of a sensor network implementing the distributed MMV BCS-GSM method.

with the first term depending on all except for the j'-th variance, while the second term depends only on the j'-th variance. This decoupling also accelerates the updating of Eq. (4.32). In particular, $l(\sigma_{j'}^{-2})$ is given by:

$$l(\sigma_{j'}^{-2}) = \frac{1}{2}\Big[\sum_{i=1}^{K}\big(\log(\sigma_{j'}^{-2}) - \log(\sigma_{j'}^{-2} + s_{i,j'}) + \frac{q_{i,j'}^2}{\sigma_{j'}^{-2} + s_{i,j'}}\big)\Big] \tag{4.34}$$

where

$$s_{i,j'} = \vec{\phi}_{i,j'}^T \mathbf{C}_{i,-j'}^{-1} \vec{\phi}_{i,j'} \quad \text{and} \quad q_{i,j'} = \vec{\phi}_{i,j'}^T \mathbf{C}_{i,-j'}^{-1} \vec{g}_i \ , \tag{4.35}$$

with $\vec{\phi}_{i,j'}$ denoting the j'-th column of $\mathbf{\Phi}_i$ and $\mathbf{C}_{i,-j'}$ is equal to $\mathbf{C}_i$ with the contribution of j'-th basis vector $\vec{\phi}_{i,j'}$ removed.

4.7.3 Estimation of a sparse vector $\vec{w}$

The fact that all covariance matrices $\{\mathbf{P}_i\}_{i=1}^K$ depend on the same model parameters $(\sigma_\eta^2, \mathbf{D})$ means that the algorithm will converge to a common set of indices $\mathscr{B}$ indicating the significant basis vectors which are used from each measurement matrix $\mathbf{\Phi}_i$. In addition, the fact that the marginal log-likelihood is written as a sum of K terms suggests that the estimation process can be distributed over the K sensors of the network, with each sensor employing the SMV BCS-GSM method described in the previous section in order to obtain a local estimation of the sparse vector $\vec{w}$. For convenience, in the subsequent analysis we describe the distributed implementation in terms of a sensor network scenario.

As we mentioned before, the proposed MMV BCS-GSM algorithm is developed under the assumption of a common sparsity structure for the signal models of the K sensors. This means that at each iteration the necessary information should be exchanged in the network, such that all sensors activate the same basis function. By exploiting the rule described in Section 4.3.2 for the addition and deletion of a basis function in the SMV case, we introduce the following heuristic rule for the addition or deletion of a basis function in the signal models of all sensors when multiple measurement vectors are available.

- ❒ the i-th sensor, $i \in \{1, \ldots, K\}$, computes the differences $\xi_{i,j} = q_{i,j}^2 - s_{i,j}$ (Eq. (4.35)) for every column $j \in \{1, \ldots, N\}$ of its measurement matrix $\mathbf{\Phi}_i$, along with the corresponding estimation of σ_j^{-2} (Eq. (4.16)). Then, if for all sensors $\xi_{i,j} > 0$ and $\sigma_j^{-2} < \infty$, each sensor re-estimates σ_j^{-2}.

- ❒ if for all sensors $i \in \{1, \ldots, K\}$ it holds that $\xi_{i,j} > 0$ and $\sigma_j^{-2} = \infty$, then the j-th basis function should be added in the corresponding model and the set of significant basis vectors is updated, $\mathscr{B} \leftarrow \mathscr{B} \cup \{j\}$. In addition, each sensor updates its associated σ_j^{-2}.

- ❒ if there is at least one sensor i' for which $\xi_{i',j} \leq 0$ and $\sigma_j^{-2} < \infty$, then the j-th basis vector is deleted from the K models (measurement matrices) and the set of significant basis vectors is updated, $(\mathscr{B} \leftarrow \mathscr{B} \setminus \{j\})$. Besides, each sensor sets $\sigma_j^{-2} = \infty$.

From the above heuristic rule we observe that the minimal information which should be exchanged between the nodes of the network in order to decide if their associated j-th basis vector should be added or deleted from the corresponding model is simply the sign of $\xi_{i,j}$ and an indicator distinguishing between the two cases $\sigma_j^{-2} = \infty$ and $\sigma_j^{-2} < \infty$.

Each iteration of the proposed algorithm results in K multivariate Gaussians with parameters $(\vec{\mu}_i, \mathbf{P}_i)$ given by Eqs. (4.30)-(4.31) by employing the model parameters estimated by maximizing Eq. (4.34). In contrast to the single measurement vector case, this objective function presents several *local maxima*. Experimental evaluation revealed that in some cases, depending on the values of $s_{i,j'}$ and $q_{i,j'}$, the numerical optimization of this expression was unstable in the sense that the estimated local maximum was far away from the global one or the algorithm oscillated among two local maxima close to each other. For this purpose, the "optimal" $\sigma_{j'}^{-2}$ can be defined by averaging the unique global maxima of the additive terms of Eq. (4.34).

Since we are working in a Bayesian framework, we are interested in combining these Gaussians in a single "best" representative, which will be then considered to be the estimation of the sparse weight vector $\vec{w}$. For this purpose, a clustering method exploiting the statistical assumptions made in this work must be used. The standard k-Means-based techniques exploit only first-order moments and they also assume that each input is a sample drawn from one of k distributions (here Gaussians). However, in the proposed method we have to incorporate the information of the second-order moments (covariance matrices), since they represent the heavy-tailed behavior of $\vec{w}$, which is crucial for achieving an increased sparsity and also we are interested in clustering a set of distributions (and not a set of samples) into a single best representative.

For this purpose, we employ a *differential entropic clustering* (DEC) of the K multivariate Gaussians [125] and the best representative $(\vec{\mu}^*, \mathbf{P}^*)$ is defined as the multivariate Gaussian that minimizes the total differential entropy with respect to the K Gaussians. Thus, at the end of each iteration the *estimated value* of $\vec{w}$ is defined as $\hat{\vec{w}} \equiv \vec{\mu}^*$, where

$$\vec{\mu}^* = \sum_{i=1}^{K} c_i \vec{\mu}_i \ , \tag{4.36}$$

$$\mathbf{P}^* = \sum_{i=1}^{K} c_i \big(\mathbf{P}_i + (\vec{\mu}_i - \vec{\mu}^*)(\vec{\mu}_i - \vec{\mu}^*)^T\big) \ , \tag{4.37}$$

where c_i are weights, which we choose to be equal to $1/K$. It is also important to note that the update of A from Eq. (4.5) is carried out by substituting $\vec{w}$ with $\vec{\mu}^*$.

Algorithm 2 summarizes the proposed BCS-GSM reconstruction approach when multiple measurement vectors are available. We use the same convergence criterion as in the standard BCS-GSM algorithm, namely, the algorithm terminates when the relative increase of the

Algorithm 2 Estimation of a sparse vector $\vec{w}$ via BCS-GSM using multiple measurement vectors

Input: $\{\mathbf{\Phi}_i\}_{i=1}^K$, $\mathbf{G}$, $c \sim 10^{-3}$

Output: $\hat{\vec{w}} \equiv \vec{\mu}^*$, $\mathbf{P}^*$, σ_η^2, $\mathscr{B}$ {the set of significant basis functions}

Initialize: $\sigma_\eta^2 = c \cdot \frac{1}{K}\sum_{i=1}^K Var(\vec{g}_i)$

select "significant" index j_1 s.t. $j_1 = \arg\max_{j=1,\ldots,N} \frac{\frac{1}{K}\sum_{i=1}^K \|\vec{\phi}_{i,j}\|^2}{\left(\frac{1}{K}\sum_{i=1}^K \|\vec{\phi}_{i,j}^T\vec{g}_i\|^2/\|\vec{\phi}_{i,j}\|^2\right)-\sigma_\eta^2}$

set $\sigma_{j_1}^{-2} = \frac{\frac{1}{K}\sum_{i=1}^K \|\vec{\phi}_{i,j_1}\|^2}{\left(\frac{1}{K}\sum_{i=1}^K \|\vec{\phi}_{i,j_1}^T\vec{g}_i\|^2/\|\vec{\phi}_{i,j_1}\|^2\right)-\sigma_\eta^2}$ (all other $\{\sigma_j^{-2}\}_{j\neq j_1}$ are set to infinity)

$\mathscr{B} = \{j_1\}$

Compute $\{\mathbf{P}_i\}_{i=1}^K$ (Eq. (4.31)), $\{\vec{\mu}_i\}_{i=1}^K$ (Eq. (4.30)) (initially scalars)
Compute $\vec{\mu}^*$ (Eq. (4.36)), $\mathbf{P}^*$ (Eq. (4.37)) using DEC and estimate A (Eq.(4.5)) by setting $\vec{w} = \vec{\mu}^*$
repeat
  for $i = 1, \ldots, K$ **do**
    for $j = 1, \ldots, N$ **do**
      Compute $\xi_{i,j} = q_{i,j}^2 - s_{i,j}$
      if $\forall\, i\ ,\ \xi_{i,j} > 0$ and $\sigma_j^{-2} < \infty$ **then**
        re-estimate σ_j^{-2}
      else if $\forall\, i\ ,\ \xi_{i,j} > 0$ and $\sigma_j^{-2} = \infty$ **then**
        add j-th basis in the model ($\mathscr{B} \leftarrow \mathscr{B} \cup \{j\}$) and update σ_j^{-2}
      else if $\exists\, i'\ ,\ \xi_{i',j} \leq 0$ and $\sigma_j^{-2} < \infty$ **then**
        delete j-th basis from the model ($\mathscr{B} \leftarrow \mathscr{B} \setminus \{j\}$) and set $\sigma_j^{-2} = \infty$
      end if
      Update $\{\mathbf{P}_i\}_{i=1}^K$, $\{\vec{\mu}_i\}_{i=1}^K$
      Update $\vec{\mu}^*$, $\mathbf{P}^*$ using DEC and estimate A
      Update $\sigma_\eta^2 = \frac{1}{K}\sum_{i=1}^K \frac{\|\vec{g}_i - \mathbf{\Phi}_i\vec{\mu}_i\|^2}{N - card(\mathscr{B}) + \sum_{n\in\mathscr{B}} A^{-1}\sigma_n^{-2}\mathbf{P}_{i,nn}}$ {*card*: cardinality of a set}
      Update $\mathbf{D}$ by performing the scaling $A\,\sigma_j^2$
    end for
  end for
until convergence

marginal log-likelihood falls below a predefined threshold for all $i \in \{1, \ldots, K\}$. Notice also that the computational complexity of the proposed approach increases by a multiplicative factor at the order of $\mathcal{O}(K)$ when compared with the complexity of the single measurement vector BCS-GSM implementation.

4.7.4 Adaptive BCS-GSM using multiple measurement vectors

The procedure described in Section 4.4 for selecting adaptively the future measurements so as to minimize the uncertainty in the estimation of $\vec{w}$ and consequently of the original signal $\vec{x}$, can be easily adopted in the case of multiple measurement vectors. What we have to do is to execute Algorithm 2 by augmenting each measurement matrix $\mathbf{\Phi}_i$ with an optimally selected next projection $\vec{\phi}_{i,M+1}$ adapted to the information content of the sparse vector $\vec{w}$ with the goal of reducing the uncertainty of $\vec{x}$. This is impossible with the previous norm-based approaches.

Following the differential entropy minimization process, the next optimal projection is carried out by performing an eigen-decomposition of $\mathbf{P}_i$ (Eq. (4.31)) and selecting $\vec{\phi}_{i,M+1}$ to be the eigenvector corresponding to the largest eigenvalue.

4.8 Distributed implementation of MMV BCS-GSM

In the following, the proposed MMV BCS-GSM method is carried out by the nodes of a sensor network in a distributed fashion using two distinct implementations corresponding to the star-shaped and the ring-shaped topologies shown in Figures 4.22(a), 4.22(b). We also consider that the k-th node possess the measurement matrix $\mathbf{\Phi}_k$, which is used for the construction of the corresponding measurement vector $\vec{g}_k$, $k = 1, \ldots, K$.

4.8.1 Distributed MMV BCS-GSM using a star-shaped topology

In a WSN scenario employing a star-shaped network topology (cf. Figure 4.22(a)) we are interested in reducing the communication cost between the nodes and the fusion center (FC) as much as possible. For this purpose, we propose an implementation where each sensor transmits the minimal amount of data required to reconstruct the signal of interest. The reconstruction process is performed in two stages alternated sequentially. In the first stage, each sensor computes the pair $(\xi_{i,j}, \sigma_j^{-2})$ and transmits to the fusion center the sign of $\xi_{i,j}$ and an indicator distinguishing between the two cases $\sigma_j^{-2} = \infty$ and $\sigma_j^{-2} < \infty$. The FC collects this information and applies the heuristic rule described in the previous section in order to decide if the j-th basis vector should be added or removed from the signal model of each sensor. In the second stage, this decision is broadcasted to the network and each sensor updates its associated model parameters corresponding to a multivariate Gaussian distribution, namely, the pair $(\vec{\mu}_i, \mathbf{P}_i)$. This pair is transmitted to the FC, which estimates the sparse vector $\vec{w}$ as the mean vector of the multivariate Gaussian distribution obtained via differential entropic clustering (Eqs. (4.36), (4.37)).

The algorithm proceeds by alternating between these two stages until convergence. Notice also that in each iteration the i-th sensor transmits to the FC a vector $\vec{\mu}_i$ of size less than N. More specifically, the size of $\vec{\mu}_i$ (respectively of $\mathbf{P}_i$) will be equal to the cardinality of the current optimal set $\mathscr{B}$, since only the components of $\vec{\mu}_i$ (respectively columns of $\mathbf{P}_i$) with their indices belonging to $\mathscr{B}$ are taken into consideration. A further reduction in the amount of data transmitted by each sensor to the FC is achieved by exploiting the symmetry of the covariance matrix $\mathbf{P}_i$.

4.8.2 Distributed MMV BCS-GSM using a ring-shaped topology

The heuristic rule introduced in the previous section for the addition and deletion of basis vectors can be implemented in a distributed way by the nodes of a network connected in a ring-shaped topology (cf. Figure 4.22(b)), without the need for the presence of a fusion center. In this case the process starts at node 1, which computes the pair $(\xi_{1,j}, \sigma_j^{-2})$ and transmits to the next sensor the sign of $\xi_{1,j}$ and an indicator distinguishing between the two cases $\sigma_j^{-2} = \infty$ and $\sigma_j^{-2} < \infty$. The next sensor computes the sign of $\xi_{2,j}$ and an indicator distinguishing between the two cases $\sigma_j^{-2} = \infty$ and $\sigma_j^{-2} < \infty$, where σ_j^{-2} is estimated using its own measurement vector $\vec{g}_2$ and matrix $\mathbf{\Phi}_2$, and compares these values with those received from the previous sensor. If they agree the process is continued with the second sensor transmitting to the next one the sign of $\xi_{2,j}$ and the indicator corresponding to σ_j^{-2}. Otherwise, if there exists a sensor i' for which $\xi_{i',j} \leq 0$ and $\sigma_j^{-2} < \infty$ then a message is transmitted to the rest of the nodes of the ring indicating that the j-th basis vector should be deleted from the model.

Thus, after one sweep of the ring in the current iteration the K-th node decides whether the j-th basis vector must be added or deleted from the model. This corresponds to the first stage of the implementation based on a star-shaped network topology. Using this knowledge the first sensor computes the pair $(\vec{\mu}_1, \mathbf{P}_1)$, which is then transmitted to the next node. The

second node estimates its own pair ($\vec{\mu}_2$, $\mathbf{P}_2$) and combines it with the received one in order to form a partial estimation of ($\vec{\mu}^*$, $\mathbf{P}^*$) via Eqs. (4.36), (4.37). Since in the current iteration we do not have the complete knowledge of $\vec{\mu}^*$ required for the computation of $\mathbf{P}^*$ we overcome this limitation by employing the $\vec{\mu}^*$ estimated in the previous iteration. Thus, in iteration t the i-th sensor ($i > 1$) computes the pair

$$\vec{\mu}_t^* = \sum_{j=1}^{i} c_j \vec{\mu}_j \ , \tag{4.38}$$

$$\mathbf{P}_t^* = \sum_{j=1}^{i} c_j \big(\mathbf{P}_j + (\vec{\mu}_j - \vec{\mu}_{t-1}^*)(\vec{\mu}_j - \vec{\mu}_{t-1}^*)^T\big) \ , \tag{4.39}$$

which is simplified using the following recursions

$$\vec{\mu}_t^* = \sum_{j=1}^{i-1} c_j \vec{\mu}_j \ + \ c_i \vec{\mu}_i \ , \tag{4.40}$$

$$\mathbf{P}_t^* = \sum_{j=1}^{i-1} c_j \big(\mathbf{P}_j + (\vec{\mu}_j - \vec{\mu}_{t-1}^*)(\vec{\mu}_j - \vec{\mu}_{t-1}^*)^T \ + \ c_i \big(\mathbf{P}_i + (\vec{\mu}_i - \vec{\mu}_{t-1}^*)(\vec{\mu}_i - \vec{\mu}_{t-1}^*)^T\big) \ , \tag{4.41}$$

where the sums $\sum_{j=1}^{i-1}(\cdot)$ in the above two equations have been computed in the previous sensor $i-1$.

We note here that the size of $\vec{\mu}^*$ may differ from one iteration to the next one, since it depends on the cardinality of $\mathscr{B}$. For this purpose, we remove or pad with a zero the j-th component of $\vec{\mu}^*$ depending on whether the j-th basis vector has been deleted or added in the model, respectively. At the end of the second sweep, which corresponds to the second stage of the implementation based on a star-shaped topology, the K-th node has access to the pair ($\vec{\mu}^*$, $\mathbf{P}^*$), which is used to trigger the first node for the next iteration. The pair of these two sweeps is repeated until the algorithm converges.

4.9 Performance evaluation: DOA estimation

Direction-of-Arrival (DOA) estimation is a classic problem in the field of signal processing due to its numerous applications, from target tracking in a military environment to the localization of a mobile user in a smart home or a museum. Among the most prominent high-resolution techniques, MUSIC [91] detects frequencies in a signal by performing an eigen-decomposition on the covariance matrix of the received signal samples. The algorithm assumes that the number of samples and frequencies are known, with an increasing accuracy as more samples are acquired, but at the cost of a high computational complexity. MVDR [92] is based on the minimization of the output power, subject to the constraint that the gain in the steering direction is unity. Classical MVDR beamforming techniques suffer from signal suppression in the presence of errors, such as the uncertainty in the look direction and array perturbations.

In addition, all traditional DOA estimation techniques acquire the source signals by sampling them at Nyquist's rate, which may result in high storage and bandwidth requirements in many modern sensing systems. In a typical scenario, a number of sensors capture signals transmitted from several sources. The received samples are then combined to estimate the position of the sources, either by exchanging them via in-network communications among sensors or by transmitting them to a fusion center (FC) with increased power and processing capabilities. In untethered sensor arrays, the amount of transmitted information must be reduced as much as

possible, since the communication cost dominates the power consumption.

A critical observation is that the problem of DOA estimation in an environment with a number of sensors and much fewer sources presents an inherent sparsity in the space-domain. If we view the monitored field as a dense grid with the sensors and sources placed on the nodes of the grid, then each sensor can be associated with a vector with all of its components being zero except for those corresponding to the nodes of the grid where the sources are placed. This justifies the potential effectiveness of CS methods in estimating the positions of the sources, while the fact that multiple observations of the same original signals (source bearings) are available at the sensor nodes motivates the use of the proposed MMV BCS-GSM approach.

Although in the present DOA estimation scenario we consider a 2-D space, the procedure is generalized to a higher dimensional space in a straightforward way. In our setting we consider a field consisting of a linear array of K sensors and L sources. Each sensor receives a superposition of the time-domain source signals, $f(t) = \sum_{l=1}^{L} f_l(t)$. Given these received signals the goal is to determine the DOA of each source. We also assume that the sensor positions are known in advance, $\{\vec{n}_i = [x_i\,, y_i]^T\}_{i=1}^{K}$. The i-th sensor receives a time-delayed and attenuated version of the superimposed source signal $f(t)$, given by:

$$w_i(t) = a \cdot f\big(t + \Delta_i(\pi_f) - (R/c)\big)\ , \tag{4.42}$$

where a is the attenuation, $\pi_f = \theta_f$ are the unknown azimuths and $\Delta_i(\pi_f)$ is the relative time-delay at the i-th sensor of the signal transmitted at θ_f. In the following, we ignore the attenuation and assume that the $\frac{R}{c}$ term is known, or constant across the array (far-field assumption), where R is the sensor-source range and c is the speed of the propagating wave in the medium.

The azimuth space is discretized by forming a finite set of angles $\mathcal{B} = \{\pi_1, \pi_2, \ldots, \pi_N\}$ where N determines the resolution. Let $\vec{b}$ denote the sparse vector, which selects elements from $\mathcal{B}$. A non-zero component $\vec{b}_j > 0$ indicates the presence of a source at an azimuth of π_j. For $L = 1$ the sparsity pattern vector $\vec{b}$ has only one non-zero entry and thus, this is the case of the highest possible sparsity.

In particular, the angle space is discretized in 180 points, which corresponds to a resolution of 1 degree. Doing so, the sparsifying transform matrices $\{\mathbf{\Psi}_i\}_{i=1}^{K}$ will be of dimension $P \times 180$ ($P \gg 180$) with their columns containing the received time-delayed signals from each *potential source location* (elements of $\mathcal{B}$). Besides, the i-th sensor is associated with a distinct random measurement matrix $\mathbf{\Phi}_i \in \mathbb{R}^{M \times P}$, where M is the number of measurements per sensor. The bearing sparsity pattern vector $\vec{b}$ is related linearly to the received signal at the i-th sensor via the expression $\vec{w}_i = \mathbf{\Psi}_i \vec{b}$. The corresponding set of measurement vectors for the K sensors in the general noisy case is given by:

$$\vec{g}_i = \mathbf{\Phi}_i \vec{w}_i + \vec{\eta}_i = \tilde{\mathbf{\Phi}}_i \vec{b} + \vec{\eta}_i \quad , \quad i = 1, \ldots, K \tag{4.43}$$

where $\tilde{\mathbf{\Phi}}_i = \mathbf{\Phi}_i \mathbf{\Psi}_i$ and $\vec{\eta}_i \in \mathbb{R}^M$ is a noise vector with unknown variance σ_η^2. In our case, the set $\{\mathbf{\Phi}_i\}_{i=1}^{K}$ contains matrices with i.i.d. standard Gaussian entries.

The reconstruction of the sparse vector $\vec{b}$ can be also recast as an SMV problem and solved with one of the many norm-based CS approaches as follows:

$$\hat{\vec{b}} = \arg\min \|\vec{b}\|_1 \quad s.t. \quad \|\mathbf{G} - \mathbf{\Phi}\mathbf{\Psi}\vec{b}\|_2 < \epsilon\ , \tag{4.44}$$

where $\mathbf{G} = [\vec{g}_1, \ldots, \vec{g}_K]$ contains the multiple measurement vectors, $\mathbf{\Psi} = [\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_K]^T$, $\mathbf{\Phi} = diag(\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_K)$ is a block-diagonal measurement matrix and ϵ is the noise level. In the following, we compare the performance of the proposed MMV BCS-GSM algorithm (M-BCS-

GSM) described in Algorithm 2, with the performance of Bayesian (BCS, BCS-GSM) and norm-based CS methods (GPSR, BP, StOMP, L1EQ-PD, SL0), which solve the above optimization problem (4.44).

In the subsequent performance evaluation the time-delay between each potential source position and the sensor is computed for $c = 340$ m/s with a sampling frequency $f_s = 500$ Hz, while the source signals are generated by drawing 512 samples from a standard Gaussian distribution $\mathcal{N}(0, 1)$. Besides, the sensor array consists of 5 sensors placed at a distance of 10^o (on the grid) from each other. The SNR at the leftmost sensor is equal to 20 dB and it reduces at 1.5 dB from sensor to sensor as we are moving on the right side of the array. We illustrate the efficiency of the proposed method for estimating DOAs in two test cases: 1) presence of a single source placed at 54^o and 2) presence of two sources with small angular separation at 41^o and 44^o, respectively, in order to evaluate the discrimination capability of the several CS reconstruction methods. In each case the results are averaged over 100 Monte-Carlo runs for a varying number of measurements per sensor, $M \in \{10, 15, 20, 25, 30\}$.

Figure 4.23(a) shows the average number of successful source detections of the source at 54^o, where a detection is characterized as successful if it recovers the positions of all sources. As we can see, the proposed method results in the highest detection performance, which increases as M increases. Its "centralized" SMV analogue (BCS-GSM) along with the norm-based methods GPSR and L1EQ-PD also perform well, but by employing more basis functions, as shown in Figure 4.23(b). This significant increase of the sparsity achieved by the proposed method is very important for resource preservation in a sensor network application.

Figure 4.24(a) shows the average number of successful source detections for the case of two sources. As in the single source case, the proposed method results in the highest detection performance for $M > 15$, with the BCS-GSM and L1EQ-PD methods following closely. However, as Figure 4.24(b) shows, the proposed method results again in an important increase of the sparsity, for instance, at the order of 80% for $M = 30$ in comparison with the BCS-GSM and at an even higher order when compared with the other methods.

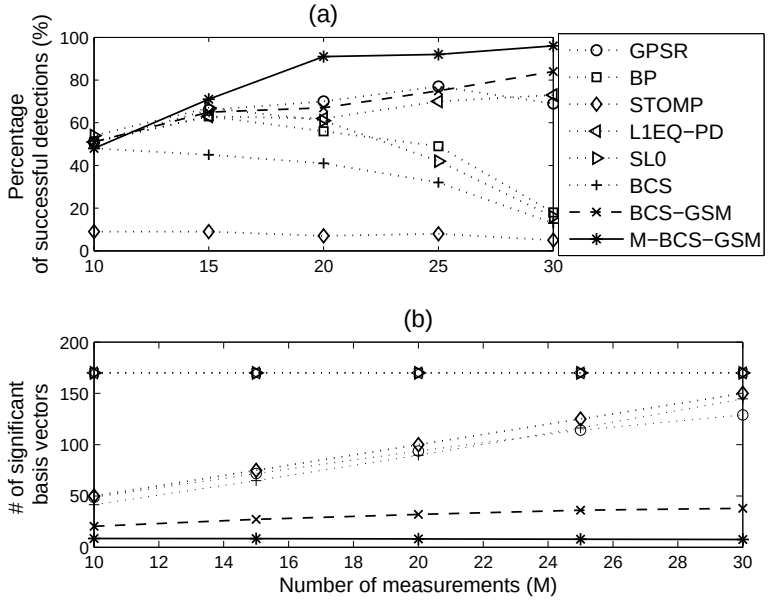


Figure 4.23: DOA estimation performance for one source (54^o).

In the following, we test the estimation performance of the proposed MMV BCS-GSM algorithm in a more general network setup for a varying number of sensors and sources, as well as

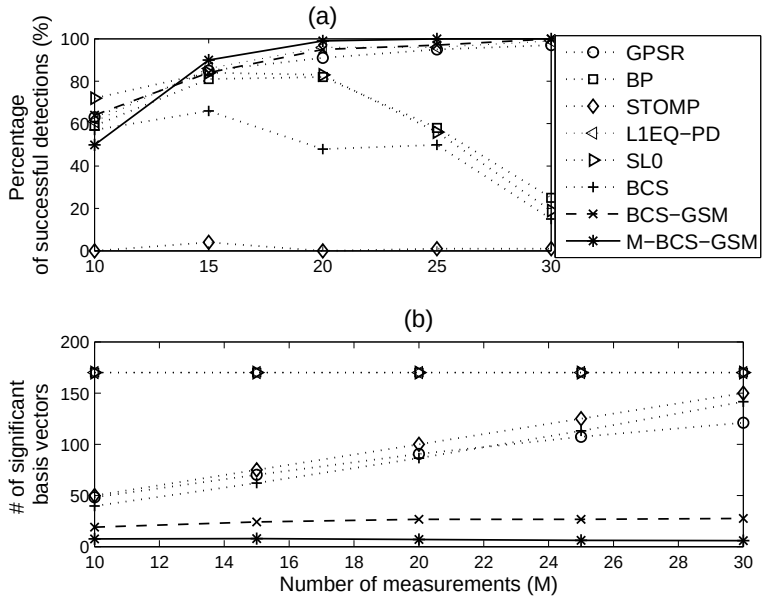


Figure 4.24: DOA estimation performance for two sources $(41^{\text{o}}, 44^{\text{o}})$.

for varying SNR values. The time-delay is computed for $c = 340$ m/s with a sampling frequency $f_s = 500$ Hz, while the source signals are generated by drawing 256 samples from a standard Gaussian distribution. Besides, the sensor array consists of K sensors, $K \in \{5 : 2 : 11\}$, placed at a distance of 10^{o} (on the grid) from each other, with each sensor generating $M = 25$ CS measurements. In addition, the SNR at the leftmost sensor varies in $\{30 : -5 : 15\}$ dB and it reduces at 1.5 dB from sensor to sensor as we are moving on the right side of the array.

First, we perform a series of 100 Monte-Carlo runs in order to evaluate the efficiency of the proposed method for estimating DOAs as the number of sensors and sources varies. We consider that the initial source is placed at 40^{o}, while from one set of 100 runs to the next the number of sources increases by one, with the new source being separated with an angle of 3^{o} from the rightmost source. The SNR at the leftmost sensor is set to 25 dB and it reduces at 1.5 dB from sensor to sensor. Motivated by the reconstruction performance presented in the above experiments we compare with GPSR, L1EQ-PD and BCS. Figure 4.25 shows the average detection percentages as a function of the number of sources. We observe that the proposed distributed implementation based on multiple measurement vectors (M-BCSGSM) maintains the high detection rates of its "centralized" counterpart (BCS-GSM), while, in general, it results in an increased detection performance when compared with the other methods for a small number of sensors and sources. This detection performance is achieved using much less basis vectors, since M-BCSGSM increases the CS ratio by as much as 7 times (for 2 sources and 7 sensors), as shown in Figure 4.26.

A justification for the decreased performance when the number of sensors/sources increases can be found by looking over *line 11* of Algorithm 2 describing the proposed basis selection rule. In particular, this rule suggests that a basis vector is deleted from the set of significant basis vectors $\mathscr{B}$ if there is at least one measurement vector (sensor) $i' \in \{1, \ldots, K\}$ whose corresponding signal model satisfies the proposed inequalities. Thus, as the number of measurement vectors (K) increases, the probability of affecting the set $\mathscr{B}$ by finding such a "violator" increases, too. A way to overcome this limitation is to relax the requirement for deleting a basis vector from $\mathscr{B}$. For instance, we may set a stricter condition by demanding from more than one sensors to satisfy the inequalities in *line 11* of Algorithm 2 in order to delete the corresponding basis

vector.

On the other hand, we observe that the CS ratios of M-BCSGSM and BCS-GSM follow inverse trajectories, with the first one decreasing while the second increases as the number of sources increases. A reason for the behavior of M-BCSGSM was stated before and has to do with the rule for adding and deleting basis vectors in $\mathscr{B}$. As the number of sources increases, the number of columns P of the measurement matrices $\mathbf{\Phi}_i$ increase, too (cf. Eq. (4.43)). This is equivalent to increasing the over-completeness of the dictionary (measurement matrix $\mathbf{\Phi}_i$) and thus the space of candidate sparse models.

In the second test case we fix the number of sensors to $K = 5$, while the number of sources varies from 2 to 6, placed as described in the previous test case, and the initial SNR at the leftmost sensor is reduced in each set of 100 Monte-Carlo realizations from 30 dB to 15 dB. Figure 4.27 shows the average detection percentages of M-BCSGSM compared with the performance of BCS-GSM, GPSR, L1EQ-PD and BCS, as a function of the number of sources. We observe that the proposed distributed implementation outperforms the other CS methods in most of the cases, even its centralized counterpart (BCS-GSM) for environments with 2 sources. Besides, the detection performance of M-BCSGSM is robust as the initial SNR decreases, while also employing a significantly reduced number of basis vectors, as shown in Figure 4.28.

The robustness of the achieved CS ratios for a decreasing SNR is due to the design of M-BCSGSM, as well as of BCS-GSM. In particular, both methods perform a correction of the noise variance in each iteration by dividing σ_η^2 with the estimated scaling factor A of the GSM model. As a result, both algorithms adapt to the true SNR value until their convergence and thus their sparsity remains unaffected in practice.

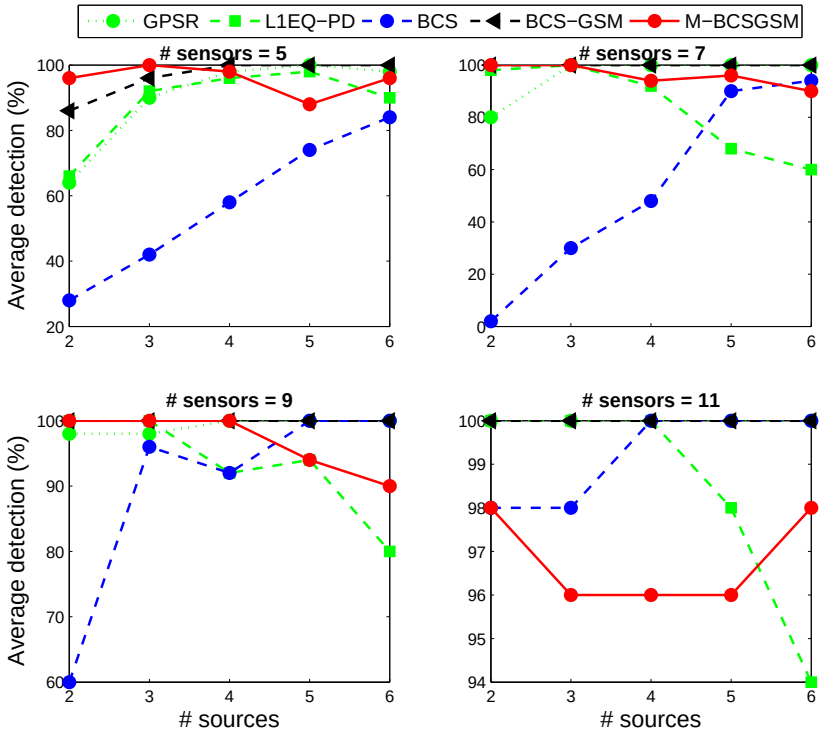


Figure 4.25: Average percentages of successful DOA estimations for a varying number of sources and sensors.

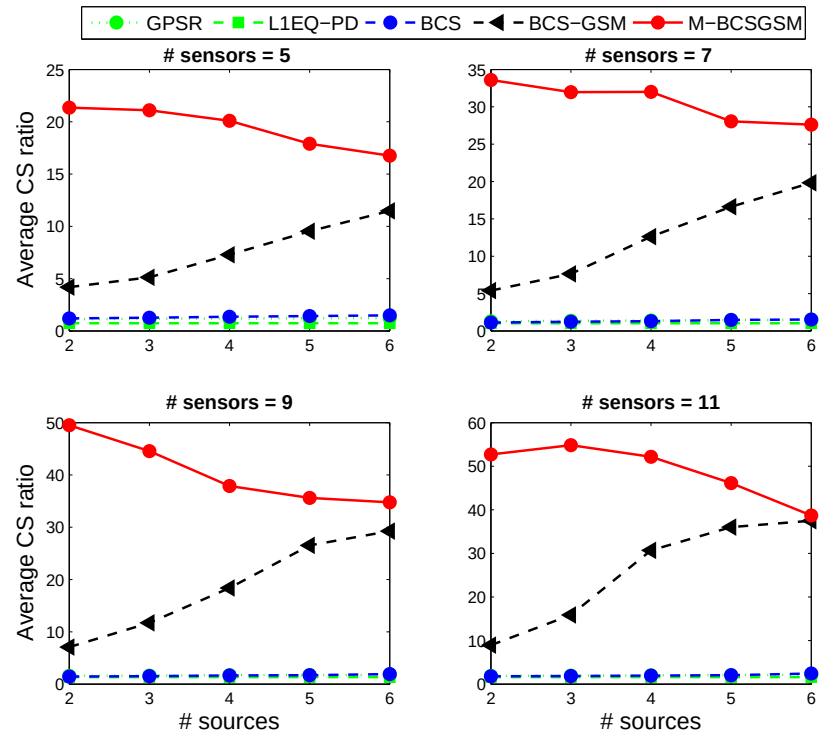


Figure 4.26: Average CS ratios for a varying number of sources and sensors.

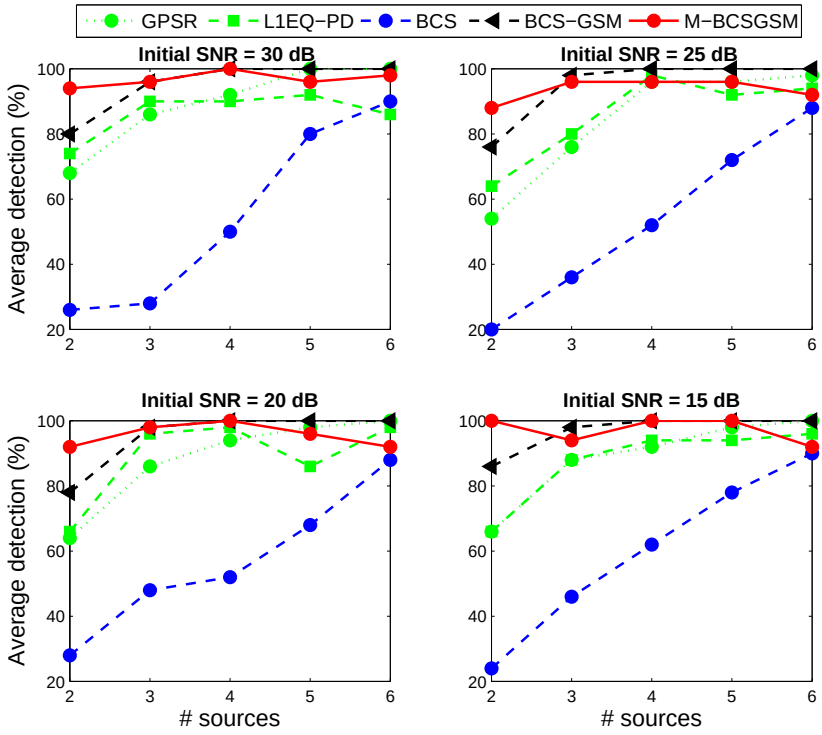


Figure 4.27: Average percentages of successful DOA estimations for a varying number of sources and initial SNR values at the leftmost sensor ($K = 5$).

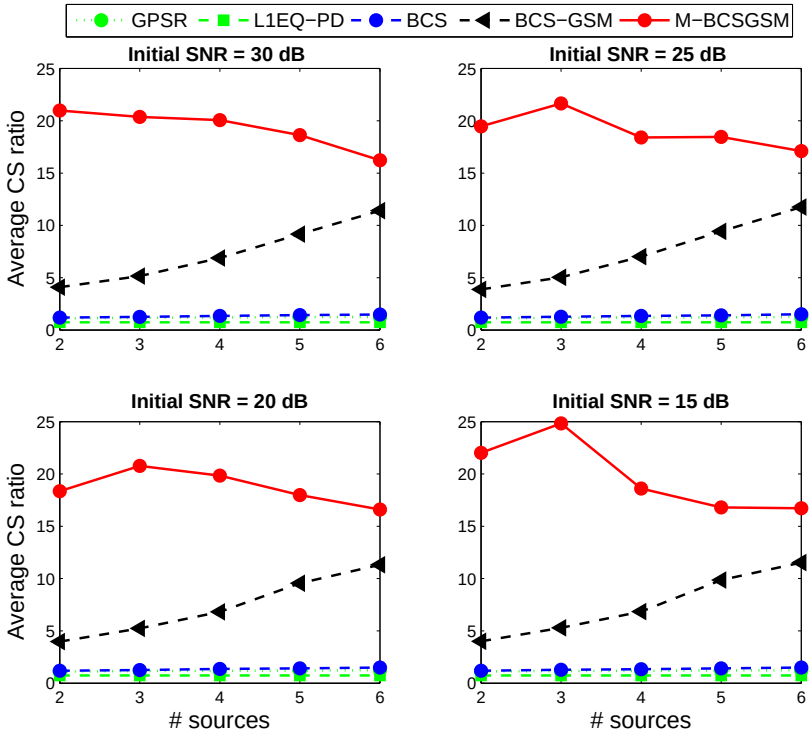


Figure 4.28: Average CS ratios for a varying number of sources and initial SNR values at the leftmost sensor ($K = 5$).

4.10 Performance evaluation: Image reconstruction

As a second illustration, we test the performance of the proposed MMV BCS-GSM (M-BCSGSM) algorithm in recovering real-world images. First, we apply Algorithm 2 on the set of medical images shown in Figure 4.13. Each image is decomposed in 5 scales via a 2-D DWT using the "db 4" wavelet. Besides, two noisy versions are generated for each image corresponding to SNR $= 7.5, 15$ dB. A hybrid reconstruction approach is followed by applying the CS algorithms on a portion of the detail coefficients only, while the approximation coefficients are employed without compression. In particular, we consider a varying number of CS measurements $M = c \cdot N_{detail}$, where N_{detail} is the number of detail coefficients and $c \in \{0.3, 0.4, 0.5, 0.65\}$.

Figures 4.29, 4.30 show the PSNRs between the reconstructed (noiseless and noisy) images and the corresponding original (noiseless) image, for the M-BCSGSM method employing K measurement vectors, $K \in \{5, 10, 15, 20\}$, as a function of the number of CS measurements M for the two SNR values. First, we observe that, in general, the reconstruction performance increases as M increases, while the decrease of SNR yields a degradation of the reconstruction quality, something that we expected. Although the differences between the PSNR values are almost negligible, we observe that an increased number of measurement vectors does not result necessarily in an increased reconstruction quality. However, we could say that as M decreases fewer measurement vectors suffice in achieving a higher PSNR.

Similar remarks can be extracted for the associated CS ratios presented in Figures 4.31, 4.32 on page 81. More specifically, we observe that the CS ratios increase as M increases and the rate of increase is robust even in the low-SNR case. However, a larger number of measurement vectors does not necessarily result in a higher CS ratio as in the case of PSNR. As in DOA estimation, a justification for this behavior can be found in *line 11* of Algorithm 2. In particular, the proposed basis selection rule suggests that a basis vector is deleted from the set of significant basis vectors $\mathscr{B}$ if there is at least one measurement vector $i' \in \{1, \ldots, K\}$ whose corresponding signal model satisfies the proposed inequalities. Thus, as K increases the probability of affecting the set $\mathscr{B}$ by finding such a "violator" increases, too. A way to overcome this limitation is to relax the requirement for deleting a basis vector from $\mathscr{B}$. For instance, we may set a stricter condition by demanding from more than one (out of K) signal models to satisfy the inequalities in *line 11* of Algorithm 2 in order to delete the corresponding basis vector.

For convenience, in Figures 4.33, 4.34 on page 82 we compare the reconstruction performance of the M-BCSGSM method with its SMV counterpart (BCS-GSM). As we can see, both methods achieve practically the same PSNR, which is closer to the optimal value achieved by the linear reconstruction (inverse 2-D DWT) as the SNR decreases. In addition, the M-BCSGSM algorithm results in an increased CS ratio especially as M decreases, as shown in Figures 4.35, 4.36 on page 83, while both methods result in a significant increase of the CS ratio when compared with the standard BCS approach.

As a last illustration, we test the reconstruction performance of M-BCSGSM on the set of images shown in Figure 4.10, with each image being decomposed in 6 scales using the "db 4" wavelet. In contrast to the medical images, the experimental results revealed an increased performance when the measurement vectors are obtained using Hadamard measurement matrices $\mathbf{\Phi}$, instead of matrices with i.i.d. standard Gaussian entries. The M-BCSGSM extension achieves the same performance when compared with BCS-GSM in terms of PSNR and CS ratios (cf. Table 4.1). Figures 4.37- 4.39 on page 85 show the original (noiseless) and noisy (for SNR $= 5$ dB) reconstructed images obtained from the optimal linear approach (inverse 2-D DWT) along with BCS, BCS-GSM and M-BCSGSM (with $K = 20$) by employing 60% of the detail coefficients. These figures reveal an interesting characteristic of M-BCSGSM when applied on this set of images with distinct content. In particular, as it can be seen from the reconstructed images corresponding to the noisy case, the M-BCSGSM algorithm results in an

increased denoising performance. Besides, it tends to smooth the edges simulating the action of a diffusion operator. This remarkable behavior requires a more thorough study before turning out in more general conclusions.

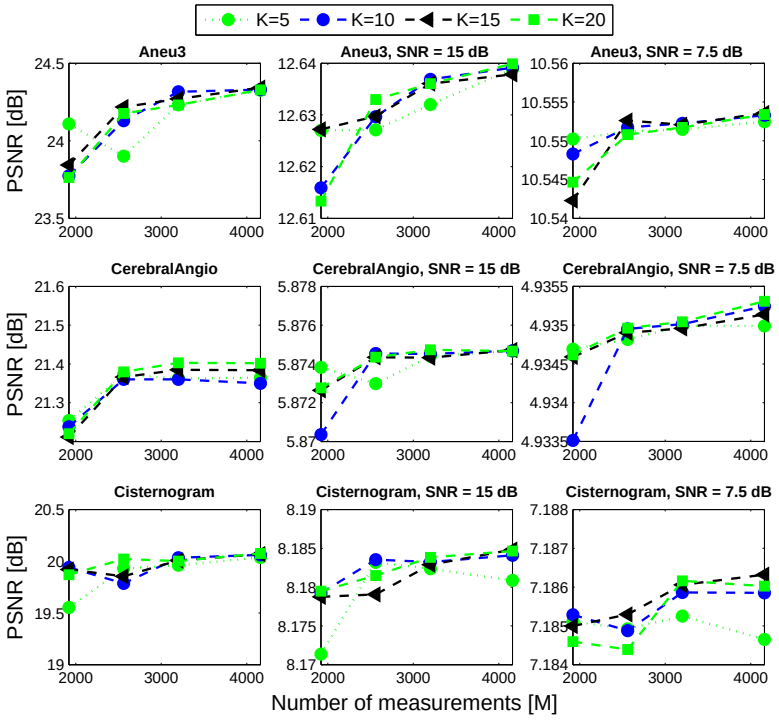


Figure 4.29: PSNRs for CS reconstruction using MMV BCS-GSM of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB).

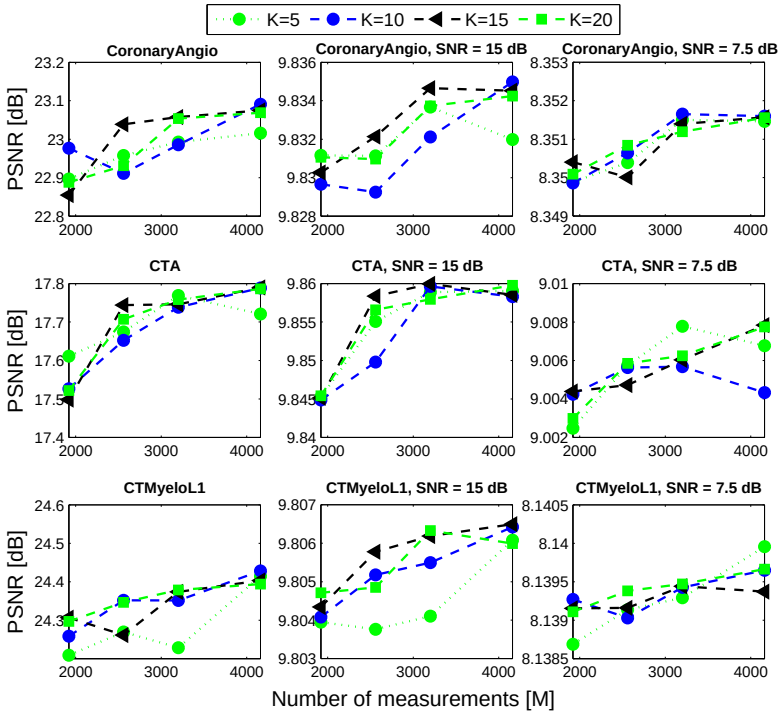


Figure 4.30: PSNRs for CS reconstruction using MMV BCS-GSM of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB).

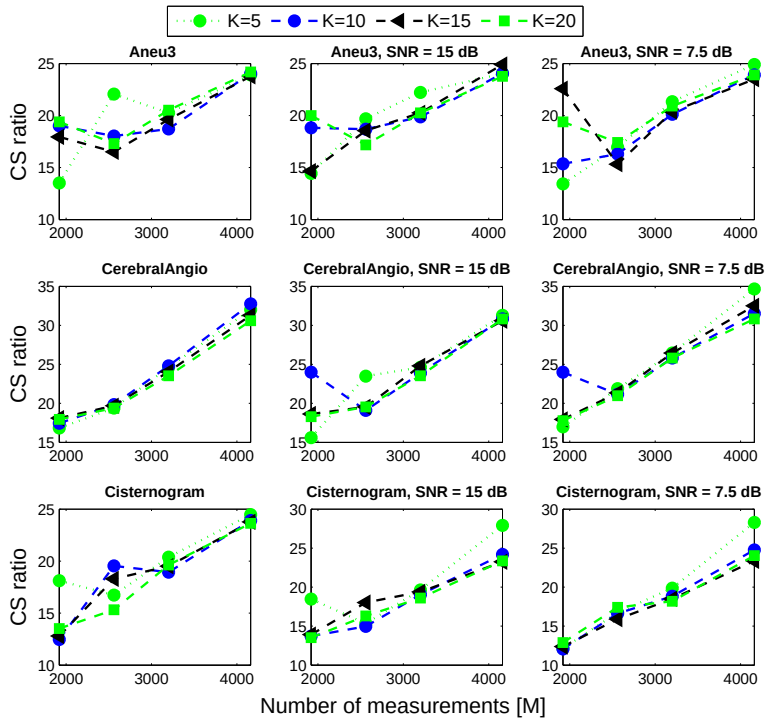


Figure 4.31: CS ratios for CS reconstruction using MMV BCS-GSM of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB).

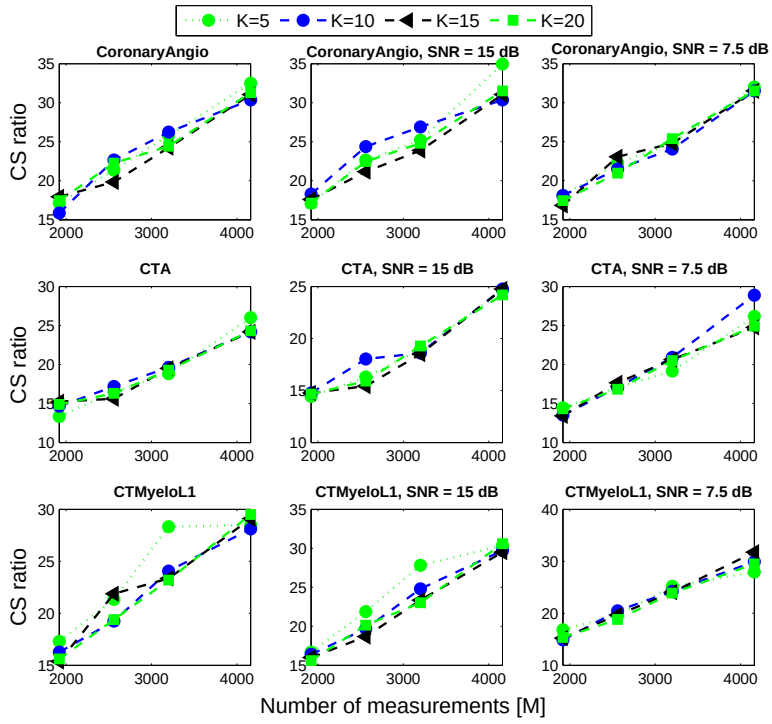


Figure 4.32: CS ratios for CS reconstruction using MMV BCS-GSM of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M, for $K \in \{5, 10, 15, 20\}$ (SNR $= 7.5, 15$ dB).

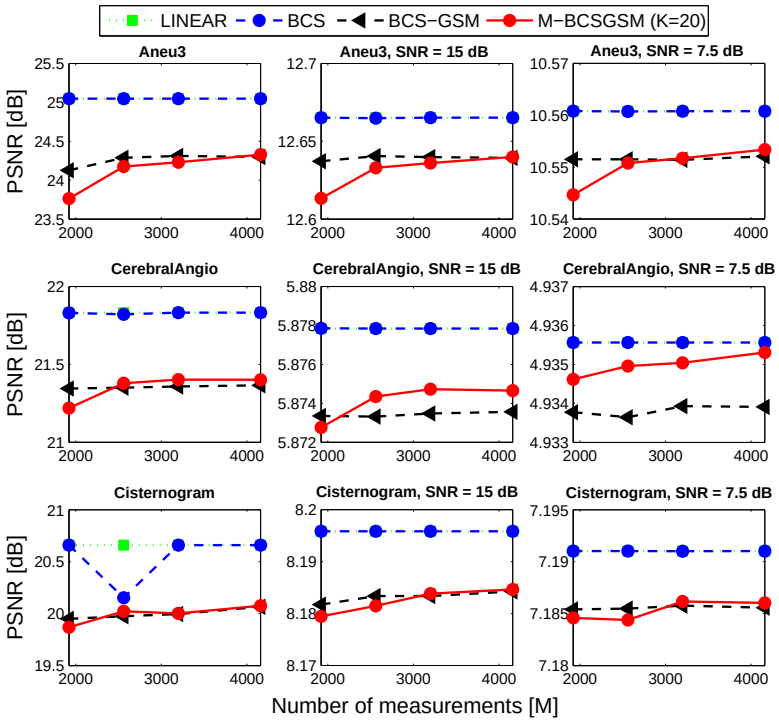


Figure 4.33: PSNRs comparison between Linear (optimal) reconstruction with BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M (SNR $= 7.5, 15$ dB).

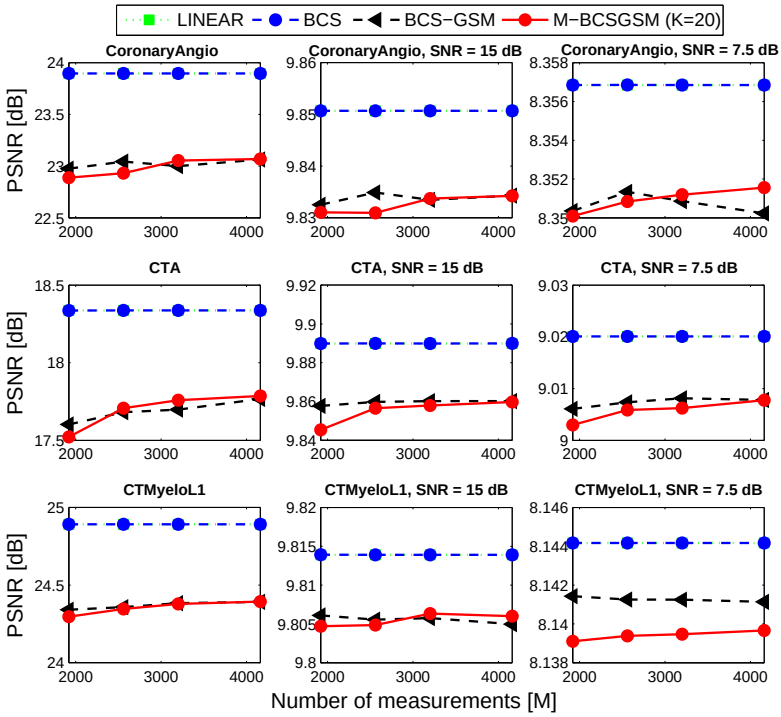


Figure 4.34: PSNRs comparison between Linear (optimal) reconstruction with BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M (SNR $= 7.5, 15$ dB).

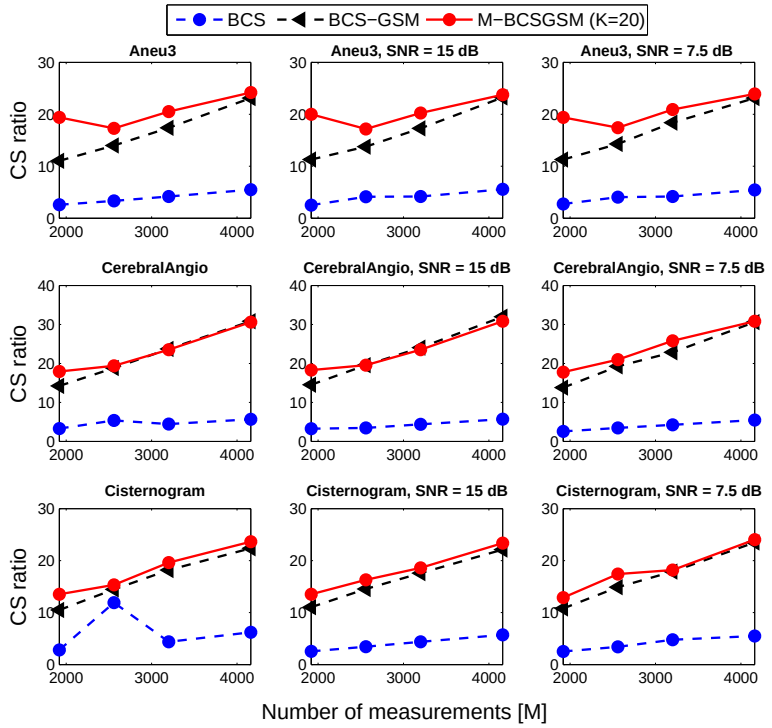


Figure 4.35: CS ratios comparison between BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "Aneu3", "CerebralAngio" and "Cisternogram" as a function of M (SNR $= 7.5, 15$ dB).

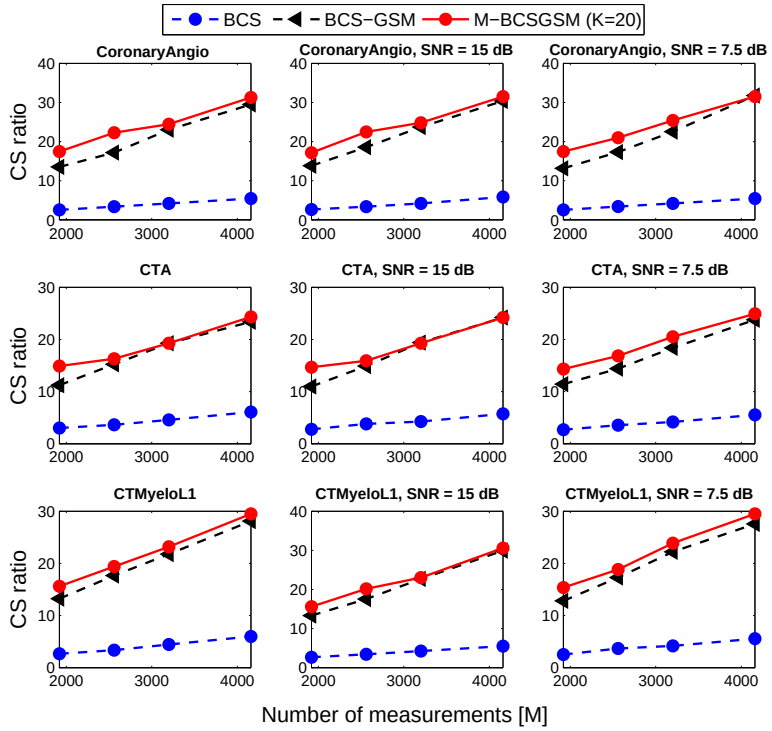


Figure 4.36: CS ratios comparison between BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) of "CoronaryAngio", "CTA" and "CTMyeloL1" as a function of M (SNR $= 7.5, 15$ dB).

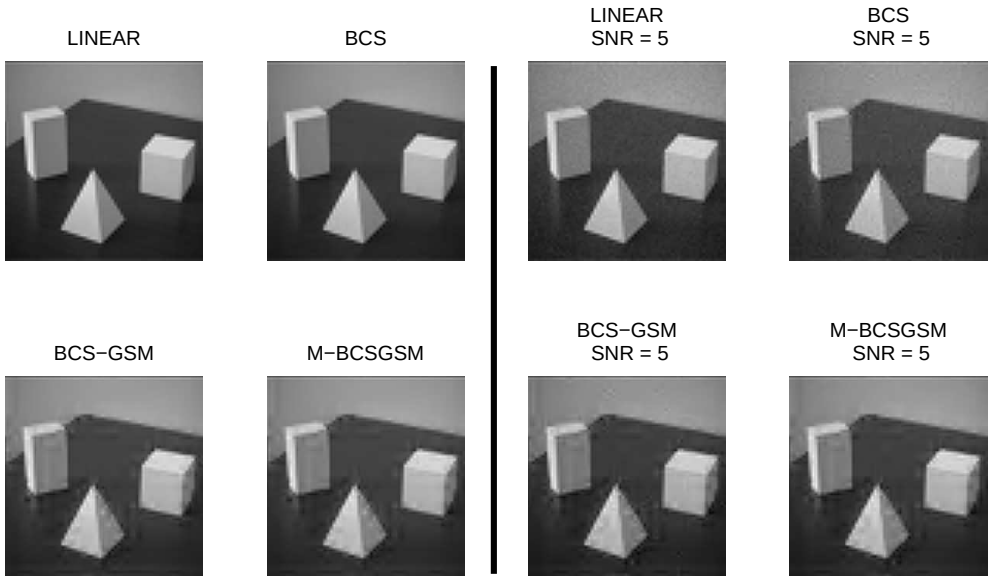


Figure 4.37: Original and noisy (SNR = 5 dB) CS reconstructed images of "Indor 2" using Linear, BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) methods.

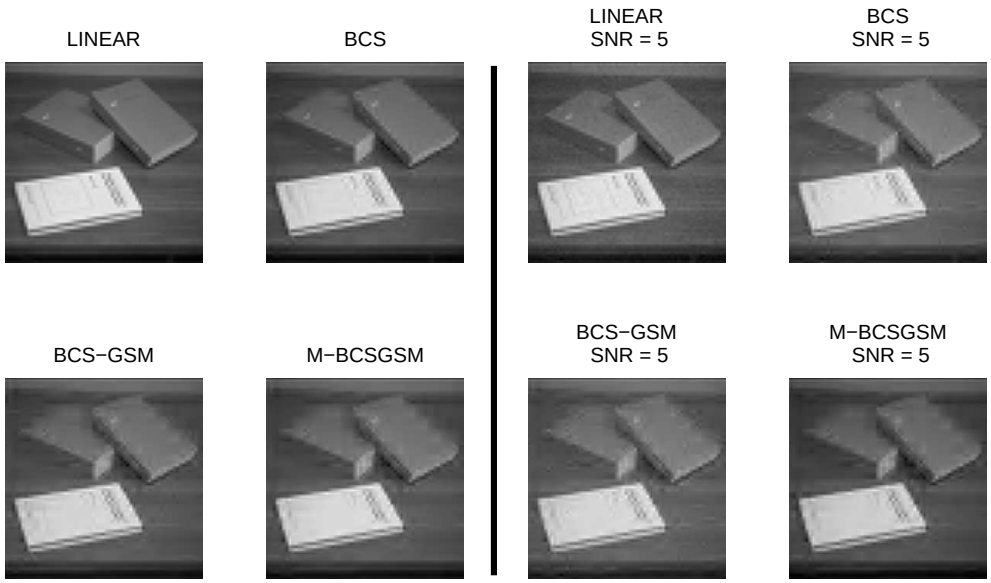


Figure 4.38: Original and noisy (SNR = 5 dB) CS reconstructed images for "Indor 4" using Linear, BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) methods.

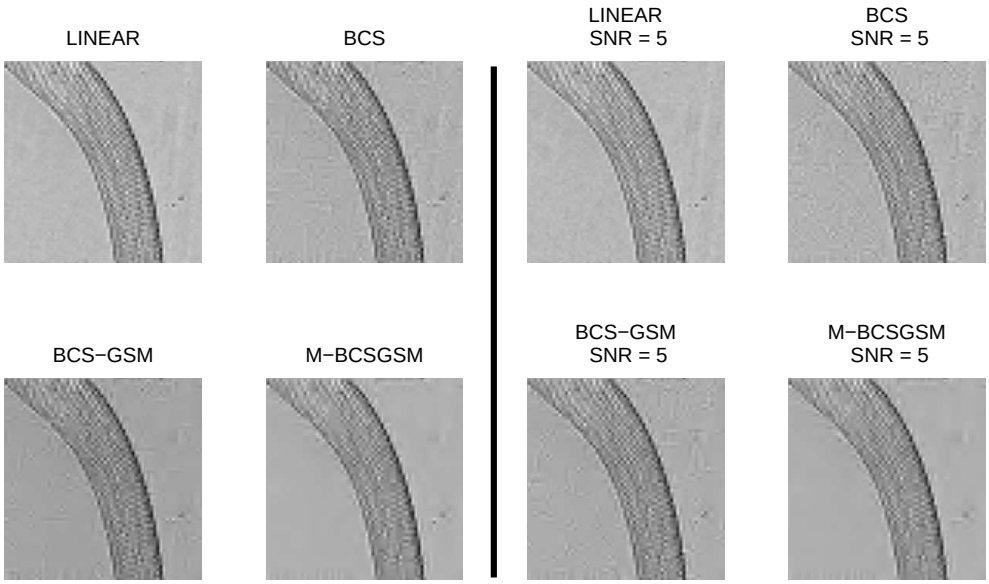


Figure 4.39: Original and noisy (SNR = 5 dB) CS reconstructed images for "Nemasup" using Linear, BCS, BCS-GSM and MMV BCS-GSM ($K = 20$) methods.

4.11 Conclusions and future work

In this chapter, we introduced a novel Bayesian CS reconstruction method. We extended a recent work [55] by replacing the hierarchical prior model with a GSM, which models directly the sparse (weight) vector with a heavy-tailed distribution that enforces its sparsity. The experimental results on 1-D synthetic signals, as well as on real-world images, revealed a critical property of the proposed BCS-GSM approach when compared with norm-based CS reconstruction methods. In particular, we showed that the BCS-GSM technique maintains comparable reconstruction performance, while using much fewer basis functions and thus resulting in an increased sparsity.

Then, the BCS-GSM algorithm was extended in the case where multiple measurement vectors are available, generated by projecting a single vector on multiple measurement matrices. The experimental results on the problems of DOA estimation and image reconstruction revealed an increased performance of the proposed MMV BCS-GSM approach when compared with the SMV BCS-GSM method, as well as with other state-of-the-art CS algorithms. In particular, we showed that the MMV BCS-GSM implementation achieves an increased performance in the problem of DOA estimation and a higher denoising capability in the case of image reconstruction, while using much fewer basis functions and thus, resulting in an increased sparsity.

In the present work we did not make any assumption for the probability density function of the scaling factor A of the GSM model. As a future work, we are interested in posing a heavy-tailed distribution on the random variable A. In particular, when A follows an α-Stable distribution (cf. Chapter 3), then the GSM is reduced to a sub-Gaussian model. We expect that the characteristic exponent which appears in the α-Stable distribution will provide further control on the sparsity of the weight vector.

Chapter 5

Bayesian Compressed Sensing of Highly Impulsive Signals in Heavy-Tailed Noise

> Great things are not done by impulse, but by a series of small things brought together.
>
> Vincent van Gogh (1853-1890)
> *Dutch Post-Impressionist artist*

5.1 Introduction

Modern acquisition devices capture signals in very high data rates and thus increasing the processing and storage requirements. According to CS theory, the complexity of a sensing system can be reduced significantly if the signal of interest is highly compressible in some orthonormal basis. Then, an accurate reconstruction can be obtained from random projections using a very small subset of the projection coefficients. A Bayesian framework was introduced in Chapter 4 with respect to the reconstruction of the original (noisy) signal, achieving an increased performance when compared with reconstruction methods employing norm-based constrained minimization approaches. The proposed BCS method was designed by employing mixtures of Gaussians to approximate the sparsity of the prior distribution of the projection coefficients.

However, there are cases in which a signal exhibits a highly impulsive behavior and thus resulting in an even sparser coefficient vector. As an illustration, Figure 5.1 on the following page shows simulated standard $S\alpha S$ sequences ($\delta = 0$, $\gamma = 1$) with increasing impulsiveness (decreasing characteristic exponent α) along with their corresponding histograms of wavelet coefficients. Each sequence is decomposed in 3 levels using "Daubechies' 4" (db 4) wavelet. It is clear that the more impulsive a signal is, the more sparse the transform coefficients are.

Highly impulsive signals are present in several natural environments, such as in underwater acoustics [126] and audio (for instance, the recordings of instruments like a drum) and consequently they are characterized by an even sparser coefficient vector. Besides, the Gaussian assumption for the statistical description of the noise is in many cases inaccurate, since there are environments where the noise (interference) distribution is heavy-tailed [127, 128, 129].

In recent BCS studies [50, 54, 55, 60], the inversion of CS measurements was considered by employing the multivariate Gaussian distribution as the prior probability model for computational purposes and in order to get closed-form expressions. In this chapter, the estimation of the sparse weight vector $\vec{w}$, and subsequently of the original impulsive signal, is also performed in a Bayesian framework. However, in order to treat the above scenarios of highly impulsive environments we develop a method which consists of modeling the prior probability of $\vec{w}$ with a heavy-tailed distribution, which promotes a high sparsity of $\vec{w}$.

The proposed reconstruction algorithm is based on a set of compressed-sensing measurements corrupted by heavy-tailed noise. The prior belief that the vector of projection coefficients should

be highly sparse is enforced by fitting its prior distribution by means of a (heavy-tailed) multivariate Cauchy distribution, which is a member of the sub-Gaussian family (cf. Section 3.2.2). In addition, a multivariate Cauchy distribution is also employed to model the heavy-tailed behavior of the noise corrupting the projection coefficients, since it models efficiently highly impulsive environments [130, 131] and also it yields closed form expressions in the subsequent Bayesian inference. The experimental results show that our proposed method achieves an improved reconstruction performance in terms of a smaller reconstruction error, while increasing the sparsity using less basis functions, when compared with a recently introduced Gaussian-based Bayesian implementation and with previous norm-based CS reconstruction algorithms.

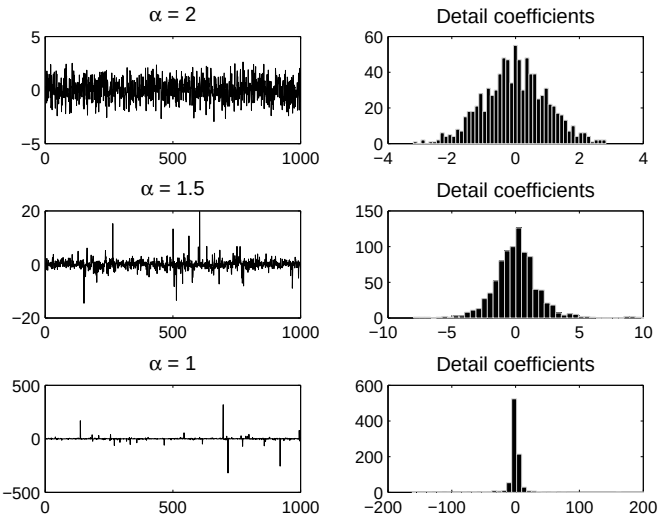


Figure 5.1: Simulated standard $S\alpha S$ sequences along with their detail wavelet coefficients histograms (3 levels, "db 4").

5.2 Statistical signal model

As we have already mentioned, the representation of the space-domain signal $\vec{x}$ is equivalent to its representation in the frequency domain $\vec{w}$. Thus, without loss of generality and similarly to the assumption we adopted in the previous chapter, in the following study we consider that the noisy CS measurements are acquired in the transform domain using the model:

$$\vec{g} = \mathbf{\Phi}\vec{w} + \vec{\eta}\,, \tag{5.1}$$

where the measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ is described in Section 1.1.1 and $\vec{\eta}$ is the associated additive noise component. Assuming that matrix $\mathbf{\Phi}$ is known, the quantities to be estimated, given the CS measurements $\vec{g}$, are the sparse weight vector $\vec{w}$ and the noise underlying variance σ_η^2. In this chapter, the assumption that $\vec{w}$ and $\vec{\eta}$ are highly sparse is formalized by modeling their prior distribution using a member of the *sub-Gaussian Symmetric α-Stable ($S\alpha S$)* family (see Definition 3.2).

In the following analysis, a multivariate sub-Gaussian distribution with underlying covariance matrix $\mathbf{\Sigma}$ (the covariance of the Gaussian part) and (in general non-zero) location parameter $\vec{\mu}$ is denoted by α-SG$(\vec{\mu}, \mathbf{\Sigma})$, where the parameter α controls the heaviness of the tails of the marginal sub-Gaussian distributions.

The multivariate Cauchy distribution (MvC) is α-SG$(\vec{\mu}, \mathbf{\Sigma})$ for $\alpha = 1$ and its density function is given by:

$$p(\vec{x}) = \frac{\Gamma\left(\frac{N+1}{2}\right)}{\pi^{(N+1)/2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \left[1 + (\vec{x} - \vec{\mu})^T \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu})\right]^{-\frac{(N+1)}{2}} , \tag{5.2}$$

where $\Gamma(\cdot)$ is the Gamma function and $|\cdot|$ denotes the matrix determinant.

In Eq. (5.1) we consider that $\vec{w} \sim$ 1-SG$(\vec{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_N^2)$ and $\vec{\eta} \sim$ 1-SG$(\vec{0}, \sigma_\eta^2 \mathbf{I}_{M\times M})$, where $\vec{0}$ is the $M \times 1$ zero vector and $\mathbf{I}_{M\times M}$ is the identity matrix. The stability property states that the sum of two sub-Gaussian (and consequently two multivariate Cauchy) distributions with the same α is again sub-Gaussian (multivariate Cauchy). Thus, the measurement vector $\vec{g} \sim$ 1-SG$(\vec{\mu}_g, \mathbf{R})$, where

$$\vec{\mu}_g = \mathbf{\Phi}\vec{\mu} , \tag{5.3}$$

$$\mathbf{R} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T + \sigma_\eta^2 \mathbf{I} . \tag{5.4}$$

From the above, it is clear that the multivariate Cauchy distribution can be viewed as a mixture of Gaussians scaled by a $S\alpha S$ random variable $A^{1/2}$, where $\mathbf{\Sigma}$ and thus $\mathbf{R}$, are determined by a discrete random vector $\vec{\tau} = [\tau_1, \ldots, \tau_N]^T$ of mixture parameters, which may be viewed as a *candidate model*. In our case it indicates which entries in $\vec{w}$ are non-zero and consequently the model is equivalent to a *basis selection*. In the following, we emphasize the dependence of the mean vector $\vec{\mu}$ (and subsequently of $\vec{\mu}_g$), as well as of the matrix $\mathbf{\Sigma}$ (and thus $\mathbf{R}$) on the mixture vector $\vec{\tau}$ by $\vec{\mu}(\vec{\tau})$ ($\vec{\mu}_g(\vec{\tau})$), $\mathbf{\Sigma}(\vec{\tau})$ ($\mathbf{R}(\vec{\tau})$), respectively. For simplicity (as in [60]), we assume that $\vec{\tau} \sim$ Bernoulli(λ_1), that is, $P\{\tau_i = 1\} = \lambda_1$ and $P\{\tau_i = 0\} = \lambda_0$. We choose $\lambda_1 \ll 1$ which ensures that, with high probability, the weight vector $\vec{w}$ has relatively few non-zero components. Accordingly, $\mathbf{\Sigma}(\vec{\tau}) = diag(\sigma_{\tau_1}^2, \ldots, \sigma_{\tau_N}^2)$ with $\sigma_{\tau_i}^2 \neq 0$ or $\sigma_{\tau_i}^2 = 0$ depending on whether the i-th component is significant and activated in the mixture or not. Similarly, $\vec{\mu}(\vec{\tau}) = (\mu_{\tau_1}, \ldots, \mu_{\tau_N})$ with $\mu_{\tau_i} \neq 0$ or $\mu_{\tau_i} = 0$. In the general case, the Gaussian part of each mixture component may be chosen from a set of Ω Gaussians with $\mu_{\tau_i} \in \{\mu_\omega\}_{\omega=1}^{\Omega}$, $\sigma_{\tau_i}^2 \in \{\sigma_\omega^2\}_{\omega=1}^{\Omega}$ and $\tau_i \in \{0, 1, \ldots, \Omega - 1\}$, where $P\{\tau_i = \omega\} = \lambda_\omega$ with $\sum_{\omega=2}^{\Omega} \lambda_\omega \ll 1$ to ensure sparsity with high probability.

In the present study, we consider the following two cases for computational simplicity:

- ❒ *Non-zero mean binary mixture*: in this case, $\Omega = 2$, $\mu_{\tau_i} \in \{0, \mu_1\}$ with $\mu_1 \neq 0$, and $\sigma_{\tau_i}^2 \in \{\sigma_0^2, \sigma_1^2\}$, where $\sigma_0^2 = 0$ to enforce sparsity. Besides, we choose $\lambda_1 \ll 1$ and $\lambda_0 = 1 - \lambda_1$.

- ❒ *Non-zero mean ternary mixture*: in this case, $\Omega = 3$, $\mu_{\tau_i} \in \{0, \mu_1, \mu_2\}$ with $\mu_1 = -\mu_2 \neq 0$, and $\sigma_{\tau_i}^2 \in \{\sigma_0^2, \sigma_1^2, \sigma_2^2\}$, where $\sigma_0^2 = 0$ and $\sigma_1^2 = \sigma_2^2$. Besides, we assume equiprobable non-zero components, that is, $\lambda_1 = \lambda_2$. This model facilitates the discrimination between active (non-zero) and non-active (zero) coefficients when no a priori knowledge of sign is available.

5.3 Estimation of a sparse vector $\vec{w}$ via an MvC prior

In this section, we describe the process for reconstructing the sparse vector $\vec{w}$ from the CS measurements $\vec{g}$. Following the above analysis, this process is reduced to finding the sparse set of the most probable basis configurations (columns of $\mathbf{\Phi}$) associated with the activated mixture components and estimating their corresponding model parameters $(\mu_{\tau_i}, \sigma_{\tau_i}^2)$. Then, their corresponding posterior probabilities are employed to obtain a Minimum Mean Squared Error (MMSE), as well as a Maximum A Posteriori (MAP) estimate of the sparse vector $\vec{w}$.

The posterior probability of a given mixture vector $\vec{\tau}'$ is given by the Bayes' rule:

$$p(\vec{\tau}'|\vec{g}) = \frac{p(\vec{g}|\vec{\tau}')p(\vec{\tau}')}{\sum_{\vec{\tau}\in\mathcal{T}} p(\vec{g}|\vec{\tau})p(\vec{\tau})} \ , \tag{5.5}$$

where in the non-zero mean binary mixture case $\mathcal{T} = \{0,1\}^N$ denotes the set containing the 2^N possible basis configurations. The posterior probabilities give a full description of the data uncertainty and are employed for inference and decision tasks. A common choice is to compute a single model that maximizes the posterior probability, the MAP estimate $\vec{\tau}_{\text{MAP}}$. However, to obtain the MMSE estimate of $\vec{w}$ we have to compute a weighted average of conditional mean estimates over all models with nonzero probability [132]. The posterior probabilities reveal the uncertainty among multiple candidate models that are ambiguous due to measurement noise or correlation among columns in the measurement matrix $\mathbf{\Phi}$.

Let $\mathcal{T}_s$ be the subset of $\mathcal{T}$ containing the mixture vectors $\vec{\tau}$ with the *most significant* posterior probabilities. We expect that the size of $\mathcal{T}_s$ will be much smaller than the size of $\mathcal{T}$ and thus, the $\{p(\vec{\tau}'|\vec{g})\}_{\vec{\tau}'\in\mathcal{T}_s}$ can be estimated from $\{p(\vec{g}|\vec{\tau}')p(\vec{\tau}')\}_{\vec{\tau}'\in\mathcal{T}_s}$. The determination of the set $\mathcal{T}_s$ requires the specification of an appropriate *selection metric*. The basis selection metric for the MvC prior model, which is used to decide whether to include a given mixture vector $\vec{\tau}$ in $\mathcal{T}_s$, or not, is defined as follows:

$$\begin{aligned}
\rho(\vec{\tau},\vec{g}) &= \ln[p(\vec{g}|\vec{\tau})p(\vec{\tau})] = \ln[p(\vec{g}|\vec{\tau})] + \ln[p(\vec{\tau})] \\
&= \ln\Big[\frac{\Gamma\big((M+1)/2\big)}{\pi^{(M+1)/2}}\Big] - \frac{1}{2}\ln\big[|\mathbf{R}(\vec{\tau})|\big] \\
&- \frac{M+1}{2}\ln\big[1 + (\vec{g}-\vec{\mu}_g(\vec{\tau}))^T\mathbf{R}(\vec{\tau})^{-1}(\vec{g}-\vec{\mu}_g(\vec{\tau}))\big] + \sum_{i=1}^{N}\ln[\lambda_{\tau_i}] \ ,
\end{aligned} \tag{5.6}$$

where

$$\sum_{i=1}^{N}\ln[\lambda_{\tau_i}] = \|\vec{\tau}\|_0 \ln\big[\lambda_1/\lambda_0\big] + N\ln[\lambda_0] \ , \tag{5.7}$$

and $\|\vec{\tau}\|_0$ is the ℓ_0-norm (the number of non-zero (activated) components) of $\vec{\tau}$.

5.3.1 MMSE and MAP estimate of $\vec{w}$

A computationally feasible approximation of the MMSE estimate of the sparse weight vector $\vec{w}$, using only the most significant posterior probabilities, is given by:

$$\hat{\vec{w}}_{\text{MMSE}} \triangleq \sum_{\vec{\tau}\in\mathcal{T}_s} p(\vec{\tau}|\vec{g})\mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}\} \ , \tag{5.8}$$

where for the approximation of $\mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}\}$ we use the underlying Gaussian part of the 1-SG$(\vec{\mu}_g,\mathbf{R})$ distribution resulting in the following expression:

$$\mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}\} = \vec{\mu}(\vec{\tau}) + \mathbf{\Sigma}(\vec{\tau})\mathbf{\Phi}^T\mathbf{R}(\vec{\tau})^{-1}(\vec{g}-\vec{\mu}_g(\vec{\tau})) \ . \tag{5.9}$$

Notice that this approximation is feasible due to the one-to-one correspondence between the probability distribution of $\vec{g}$ and its Gaussian part (Corollary 3.1). Similarly, the approximate MMSE estimation error is given by the trace of the conditional covariance matrix,

$$tr(Cov\{\vec{w}|\vec{g}\}) = \sum_{\vec{\tau}\in\mathcal{T}_s} p(\vec{\tau}|\vec{g})\Big[tr(Cov\{\vec{w}|\vec{g},\vec{\tau}\})\|\hat{\vec{w}}_{\text{MMSE}} - \mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}\}\|^2\Big] \ , \tag{5.10}$$

where

$$Cov\{\vec{w}|\vec{g},\vec{\tau}\} = \mathbf{\Sigma}(\vec{\tau}) - \mathbf{\Sigma}(\vec{\tau})\mathbf{\Phi}^T\mathbf{R}(\vec{\tau})^{-1}\mathbf{\Phi}\mathbf{\Sigma}(\vec{\tau}) \ . \tag{5.11}$$

On the other hand, the MAP basis configuration is given by $\vec{\tau}_{\rm MAP} = \arg\max_{\vec{\tau}\in\mathcal{T}_s} p(\vec{\tau}|\vec{g})$, resulting in the following approximation of the MAP estimate of $\vec{w}$:

$$\hat{\vec{w}}_{\rm MAP} \triangleq \mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}_{\rm MAP}\} \ , \tag{5.12}$$

where $\mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}_{\rm MAP}\}$ is given by Eq. (5.9) by substituting $\vec{\tau}_{\rm MAP}$ in place of $\vec{\tau}$. Thus, the MAP estimate of $\vec{w}$ is expressed only in terms of one mixture vector, namely, the vector with the highest posterior probability. This estimate could be employed in sensing systems with very limited resources, such as power, memory capacity, and bandwidth.

5.3.2 Incremental basis selection via a tree-structure

In this section, we focus our attention in extracting the set of significant mixture vectors $\mathcal{T}_s$. By observing that the joint distribution of $(\vec{\tau},\vec{g})$ is related with the basis selection metric,

$$p(\vec{\tau},\vec{g}) = p(\vec{g}|\vec{\tau})p(\vec{\tau}) = \exp\{\rho(\vec{\tau},\vec{g})\} \ ,$$

we conclude that significant values of $p(\vec{\tau},\vec{g})$ correspond to relatively large values of $\rho(\vec{\tau},\vec{g})$. In order to quantify what constitutes a relatively large value of the basis selection metric we derive its prior distribution in terms of the first two moments, namely, the mean and the variance.

If $\vec{\tau}$ is the true mixture vector from which the CS measurement vector $\vec{g}$ is generated, then

$$\begin{aligned}\mathbb{E}\{\rho(\vec{\tau},\vec{g})\} &= \ln\Big[\frac{\Gamma\big((M+1)/2\big)}{\pi^{(M+1)/2}}\Big] - \frac{1}{2}\underbrace{\mathbb{E}\Big\{\ln\big[|\mathbf{R}(\vec{\tau})|\big]\Big\}}_{\rm (I)} \\ &\quad - \frac{M+1}{2}\underbrace{\mathbb{E}\Big\{\ln\big[1+(\vec{g}-\vec{\mu}_g(\vec{\tau}))^T\mathbf{R}(\vec{\tau})^{-1}(\vec{g}-\vec{\mu}_g(\vec{\tau}))\big]\Big\}}_{\rm (II)} + \underbrace{\mathbb{E}\Big\{\sum_{i=1}^{N}\ln[\lambda_{\tau_i}]\Big\}}_{\rm (III)} \end{aligned} \tag{5.13}$$

In the following, we compute the quantities (I), (II), (III) by assuming that $\sigma^2_{\tau_i} = \sigma^2_1$ and $\lambda_{\tau_i} = \lambda_1$ for $\tau_i \neq 0$.

$$\begin{aligned}\text{(I)} &= \mathbb{E}\Big\{\ln\big[|\mathbf{R}(\vec{\tau})|\big]\Big\} = \mathbb{E}\Big\{\ln\big[(\sigma_1^2+\sigma_\eta^2)^L\sigma_\eta^{2(M-L)}\big]\Big\} \\ &= \mathbb{E}\Big\{L\ln\Big[\frac{\sigma_1^2}{\sigma_\eta^2}+1\Big] + M\ln[\sigma_\eta^2]\Big\} \\ &= \ln\Big[\frac{\sigma_1^2}{\sigma_\eta^2}+1\Big]\mathbb{E}\{L\} + M\ln[\sigma_\eta^2] \\ &\overset{L\sim\text{Binomial}}{=} \ln\Big[\frac{\sigma_1^2}{\sigma_\eta^2}+1\Big]N(1-\lambda_0) + M\ln[\sigma_\eta^2] \ . \end{aligned} \tag{5.14}$$

The second expectation of Eq. (5.13) is evaluated as follows:

$$\text{(II)} \overset{\vec{v}=\vec{g}-\vec{\mu}_g(\vec{\tau})}{=} \mathbb{E}\Big\{\ln\big[1+\vec{v}^T\mathbf{R}(\vec{\tau})^{-1}\vec{v}\big]\Big\} = \mathbb{E}\Big\{\ln\big[1+V\big]\Big\} \ , \tag{5.15}$$

where the random variable $V = \vec{v}^T\mathbf{R}(\vec{\tau})^{-1}\vec{v}$ follows a Chi-square distribution ($V \sim \chi^2_M$). The evaluation of the expectation in Eq. (5.15) falls in the general case of computing the expected

logarithm of a noncentral Chi-square random variable [133]. We estimated the above expectation by evaluating numerically the following integral using the adaptive Gauss-Kronrod quadrature method, which may be most efficient for high accuracies:

$$\int_0^\infty \ln(x+1)\frac{x^{(M-2)/2}e^{-x/2}}{2^{M/2}\Gamma\left(\frac{M}{2}\right)}\,\mathrm{d}x\ .$$

Finally, the third expectation of Eq. (5.13) is given by,

$$(\mathrm{III}) = N\ln[\lambda_0] - N(1-\lambda_0)\ln\left[\frac{\lambda_0}{\lambda_1}\right]\ . \tag{5.16}$$

Similarly, the second-order moment of $\rho(\vec{\tau},\vec{g})$ is given by:

$$Var\{\rho(\vec{\tau},\vec{g})\} = N(1-\lambda_0)\lambda_0\left(\frac{1}{4}\ln\left[\left(\frac{\sigma_1^2}{\sigma_\eta^2}+1\right)^2\right]+\ln\left[\left(\frac{\lambda_0}{\lambda_1}\right)^2\right]\right)+\frac{M+1}{2}Var\big\{\ln[1+V]\big\}\ , \tag{5.17}$$

where, analogously to the first-order case, the term $Var\big\{\ln[1+V]\big\}$ is estimated numerically by evaluating the following integral using the adaptive Gauss-Kronrod quadrature method:

$$\int_0^\infty \ln\left[(x+1)^2\right]\frac{x^{(M-2)/2}e^{-x/2}}{2^{M/2}\Gamma\left(\frac{M}{2}\right)}\,\mathrm{d}x\ .$$

Obviously, it is impractical to evaluate the posterior probabilities for all 2^N (non-zero mean binary mixture) or 3^N mixture vectors (non-zero mean ternary mixture) and in general for all Ω^N basis configurations. For this purpose, we employ a commonly used incremental tree-structured procedure for selecting the next significant basis configuration adapted to the MvC prior model. Although for convenience we focus on the binary case ($\Omega = 2$) however, the procedure is extended in a straightforward way to the general case $\Omega > 2$. The root of the tree consists of the zero vector $\vec{\tau} = \vec{0}$ for which

$$\rho(\vec{0},\vec{g}) = \ln\left[\Gamma\Big(\frac{M+1}{2}\Big)\right] - \frac{M}{2}\ln[\sigma_\eta^2] - \frac{(M+1)}{2}\ln\left[\pi\Big(1+\frac{\|\vec{g}\|^2}{\sigma_\eta^2}\Big)\right]\ . \tag{5.18}$$

At the first level of the tree the set $\mathcal{T}_1$ is formed, which contains the N binary vectors (in general the $(\Omega-1)N$ mixture vectors) $\vec{\tau}$ generated by "activating" one mixture parameter at a time. Then, the values of the selection metric $\rho(\vec{\tau},\vec{g})$ are computed for these mixture vectors and those with the K largest values are chosen to explore further and stored in $\mathcal{T}_{s,1}$, where K is a predetermined positive integer. At the second level of the tree, for each element of $\mathcal{T}_{s,1}$ a second mixture component is "activated" in all possible locations once at a time resulting in $\sum_{k=1}^{K}(N-k)$ binary vectors, which form the set $\mathcal{T}_2$. As before, the values of $\rho(\vec{\tau},\vec{g})$ are computed for these mixture vectors, and those with the K largest values are chosen to explore further and stored in $\mathcal{T}_{s,2}$. The process is repeated until the value of $\rho(\vec{\tau},\vec{g})$ is adequately large or until $\mathcal{T}_{s,l_{max}}$ is formed, where l_{max} is a predefined maximum number of tree levels, chosen such that $P\{\|\vec{\tau}\|_0 > l_{max}\}$ is sufficiently small. The explored vectors $\vec{\tau}$ that yield the significant values of $\rho(\vec{\tau},\vec{g})$ and thus of $\exp\{\rho(\vec{\tau},\vec{g})\} = p(\vec{g}|\vec{\tau})p(\vec{\tau})$ constitute the final optimal set of mixture vectors $\mathcal{T}_s$.

When moving from one level of the tree to the next one the values of the metric $\rho(\cdot)$ must be updated. In particular, the change results from the activation of a single mixture component at a time. For this purpose, let $\vec{\tau}_q$ denote the mixture vector which is identical to $\vec{\tau}$ except for the q-th component, which is "activated" in $\vec{\tau}_q$, while it is "inactive" in $\vec{\tau}$. We are interested in

computing the differences

$$\Delta_q(\vec{\tau}) = \rho(\vec{\tau}_q, \vec{g}) - \rho(\vec{\tau}, \vec{g}) \ , \tag{5.19}$$

which are then used to decide which mixture components will be activated. More specifically, the set $\mathcal{T}_{s,l}$ at the l-th tree-level is formed by keeping the K binary vectors of the set $\mathcal{T}_l$ that correspond to the K largest values of $\Delta_q(\vec{\tau})$, that is, we keep only these vectors which achieve the highest increase of the basis selection metric. From Eq. (5.6) we can see that the key quantities to be updated are the inverse of $\mathbf{R}(\vec{\tau})$ and its determinant. The update of $\mathbf{R}(\vec{\tau})$ when the q-th component is activated is given by:

$$\mathbf{R}(\vec{\tau}_q) = \mathbf{R}(\vec{\tau}) + \sigma^2_{\tau_q} \vec{\phi}_q \vec{\phi}_q^T \ , \tag{5.20}$$

and thus the matrix inversion lemma results in a simple expression for updating the inverse of $\mathbf{R}(\vec{\tau}_q)$:

$$\mathbf{R}(\vec{\tau}_q)^{-1} = \mathbf{R}(\vec{\tau})^{-1} - \gamma_q \vec{v}_q \vec{v}_q^T \ , \tag{5.21}$$

$$\gamma_q = \sigma^2_{\tau_q}(1 + \sigma^2_{\tau_q} \vec{\phi}_q^T \vec{v}_q)^{-1} \ , \tag{5.22}$$

$$\vec{v}_q = \mathbf{R}(\vec{\tau})^{-1} \vec{\phi}_q \ . \tag{5.23}$$

Besides, from Eq. (5.20) the determinant of $\mathbf{R}(\vec{\tau})$ can be easily updated as follows:

$$|\mathbf{R}(\vec{\tau}_q)| = (1 + \sigma^2_{\tau_q} \vec{\phi}_q^T \mathbf{R}(\vec{\tau})^{-1} \vec{\phi}_q) |\mathbf{R}(\vec{\tau})| = \frac{\sigma^2_{\tau_q}}{\gamma_q} |\mathbf{R}(\vec{\tau})| \ . \tag{5.24}$$

Notice also that the updated mean vector $\vec{\mu}(\vec{\tau}_q)$ (and consequently $\vec{\mu}_g(\vec{\tau}_q)$) is the same as $\vec{\mu}(\vec{\tau})$ ($\vec{\mu}_g(\vec{\tau})$) except for a change of its q-th component from $\mu_q = \mu_0 = 0$ to $\mu_q = \mu_1$. Finally, the probability of $\vec{\tau}$ is updated as:

$$p(\vec{\tau}_q) = \frac{\lambda_1}{\lambda_0} p(\vec{\tau}) \ . \tag{5.25}$$

The substitution of the above update equations in Eq. (5.6) results after some algebraic manipulation in the following expression for $\Delta_q(\vec{\tau})$ corresponding to the *MvC prior model*:

$$\Delta_q(\vec{\tau}) = \frac{1}{2} \ln\Big(\frac{\gamma_q}{\sigma^2_{\tau_q}}\Big) + \ln\Big(\frac{\lambda_1}{\lambda_0}\Big) - \frac{M+1}{2} \left(\ln\Big(1 - \frac{\gamma_q \big|\vec{\zeta}_g(\vec{\tau})^T \vec{v}_q + (\mu_1/\sigma^2_{\tau_q})\big|^2 - (\mu_1^2/\sigma^2_{\tau_q})}{1 + \vec{\zeta}_g(\vec{\tau})^T \mathbf{R}(\vec{\tau})^{-1} \vec{\zeta}_g(\vec{\tau})} \Big) \right) \ , \tag{5.26}$$

where $\vec{\zeta}_g(\vec{\tau}) = \vec{g} - \vec{\mu}_g(\vec{\tau})$. Due to our assumption that the mixture vectors consist of two components, the variances $\sigma^2_{\tau_q}$ in Eq. (5.26) can be substituted by σ_1^2. Notice that in the general case, $\Omega > 2$, we would be interested in computing the differences

$$\Delta_{q,\omega'}(\vec{\tau}) = \rho(\vec{\tau}_q, \vec{g}) - \rho(\vec{\tau}, \vec{g}) \ , \tag{5.27}$$

which quantify the change to $\rho(\vec{\tau}, \vec{g})$ that results from changing the q-th component of $\vec{\tau}$ from ω to ω'. Accordingly, in Eq. (5.26) the quantities μ_1 and $\sigma^2_{\tau_q}$ will be replaced by $\mu_{\omega',\omega} = \mu_{\omega'} - \mu_{\omega}$ and $\sigma^2_{\omega',\omega} = \sigma^2_{\omega'} - \sigma^2_{\omega}$, respectively.

In our implementation, the algorithm terminates when a total of l_{max} mixture components are "activated", that is, when we reach the l_{max}-th level of the tree. If at least one of the l_{max} values of the metric $\rho(\cdot)$ exceeds some predefined threshold ρ_{Th} the algorithm terminates. If not, a second search starts from the root and directed to ignore all previously visited nodes. This process is repeated until the threshold ρ_{Th} is exceeded or until the number of searches reaches a maximum value S_{max}. Algorithm 3 summarizes the process for estimating a highly

Algorithm 3 Estimation of a sparse vector $\vec{w}$ via an MvC prior model

Input: $\mathbf{\Phi}$, $\vec{g}$, μ_1, σ_1^2, σ_η^2, λ_1, ρ_{Th}, S_{max}, l_{max}

Output: $\hat{\vec{w}} \equiv \mathbb{E}\{\vec{w}|\vec{g}\}$, $\mathcal{T}_s$, $\{p(\vec{\tau}|\vec{g})\}_{\vec{\tau}\in\mathcal{T}_s}$, $\{\mathbb{E}\{\vec{w}|\vec{g},\vec{\tau}\}\}_{\vec{\tau}\in\mathcal{T}_s}$ (Eq. (5.9)), $\{Cov\{\vec{w}|\vec{g},\vec{\tau}\}\}_{\vec{\tau}\in\mathcal{T}_s}$ (Eq. (5.11))

Initialize: Compute $\rho(\vec{0},\vec{g})$ (Eq. (5.18))

for $i = 1,\ldots,N$ **do**
  Compute $\vec{v}_i^{(0)}$ (Eq. (5.23))
  Compute $\gamma_i^{(0)}$ (Eq. (5.22))
  Compute $\rho^{(0)}(\vec{\tau}_i,\vec{g}) = \rho(\vec{0},\vec{g}) + \Delta_i(\vec{\tau})$ (Eq. (5.19))
end for
for $s = 1,\ldots,S_{max}$ **do**
  idx = {}
  $\vec{\tau}^{(s,0)} = \vec{0}$
  for $l = 1,\ldots,l_{max}$ **do**
    $i_{opt} = \arg\max_i \rho(\vec{\tau}_i,\vec{g})$ {which also leads to an unvisited node}
    $\rho^{(s,l)} = \rho(\vec{\tau}_{i_{opt}},\vec{g})$
    $\vec{\tau}^{(s,l)} = \vec{\tau}^{(s,l-1)} + \delta_{i_{opt}}$ {$\delta_k = 1$ in k-th position and 0 elsewhere}
    idx $\leftarrow$ {idx, i_{opt}}
    for $i = 1,\ldots,N$ **do**
      Update $\vec{v}_i$
      Update γ_i
      Update selection metric: $\rho(\vec{\tau}_i,\vec{g}) = \rho^{(s,l)} + \Delta_i(\vec{\tau})$
    end for
    Estimate $\hat{\vec{w}}^{(s,l)}$ (Eq. (5.8))
    Estimate $Cov\{\vec{w}|\vec{g},\vec{\tau}\}^{(s,l)}$ (Eq. (5.11))
  end for
  if $\max\{\rho^{(s,l)}\}_{l=1,\ldots,l_{max}} > \rho_{Th}$ **then**
    terminate
  end if
end for

sparse weight vector $\vec{w}$ using an MvC prior model. For convenience, we cite the steps for the binary case.

5.4 Performance evaluation

In this section, we compare the performance of the proposed reconstruction scheme with the FBMP method proposed in [60][1]. For this purpose, we generate simulated measurement vectors $\vec{g}$ according to Eq. (5.1), where the sparse vectors $\vec{w}$ are drawn from an MvC distribution of length $N = 400$ that contain L spikes, whose locations are chosen at random. We set the sparsity as a function of λ_1 and N, $L = \lceil \lambda_1 \cdot N \rceil$. In the subsequent experiments we choose $\lambda_1 = 0.02$, which results in a highly impulsive and thus heavy-tailed vector $\vec{w}$. The measurement noise is generated by drawing samples from a zero-mean MvC distribution with underlying variance σ_η^2. The $M \times N$ measurement matrix $\mathbf{\Phi}$ is constructed by first drawing i.i.d. samples from a standard Gaussian distribution, and then by normalizing its columns to unit magnitude.

The reconstruction performance is tested for two distinct Signal-to-Noise Ratio (SNR) values

[1] We used the FBMP package downloaded from `http://www.ece.osu.edu/~zinielj/fbmp`, using the standard implementation without parameter re-estimation.

(SNR=10, 15 dB), as well as for a range of measurements ($M \in \{90 : 1 : 120\}$). In particular, the noise variance σ_η^2 and the mixture variance σ_1^2 are related via the expression:

$$SNR = \frac{\sigma_1^2 \lambda_1 N}{\sigma_\eta^2 M} \ . \tag{5.28}$$

The process is repeated for 100 independent Monte-Carlo realizations and the results are given by averaging over the 100 runs. The normalized mean-squared error of the MMSE estimated sparse vector $\hat{\vec{w}}$ is given by

$$\text{NMSE}_{\text{MMSE}} = \frac{1}{100} \sum_{j=1}^{100} \frac{\|\hat{\vec{w}}_{\text{MMSE,j}} - \vec{w}\|_2^2}{\|\vec{w}\|_2^2} \ , \tag{5.29}$$

where $\hat{\vec{w}}_{\text{MMSE,j}}$ is the MMSE estimate of $\vec{w}$, given by Eq. (5.8), at the j-th Monte-Carlo run. Similarly, the normalized mean-squared error of the MAP estimated sparse vector is given by

$$\text{NMSE}_{\text{MAP}} = \frac{1}{100} \sum_{j=1}^{100} \frac{\|\hat{\vec{w}}_{\text{MAP,j}} - \vec{w}\|_2^2}{\|\vec{w}\|_2^2} \ , \tag{5.30}$$

where $\hat{\vec{w}}_{\text{MAP,j}}$ is the MAP estimate of $\vec{w}$, given by Eq. (5.12), at the j-th Monte-Carlo run.

Figure 5.2 shows the MMSE and MAP reconstruction errors averaged over the 100 runs for SNR=10, 15 dB. It is clear that the proposed algorithm, which is based on a heavy-tailed distribution, achieves a better reconstruction performance when compared with the FBMP method, which is based on a normality assumption with respect to the prior distributions. For both methods, the corresponding MMSE and MAP estimates are close to each other. Also, the SNR level affects more the FBMP method. Besides, an interesting observation is that in a highly impulsive scenario the increasing number of measurements deteriorates the reconstruction performance.

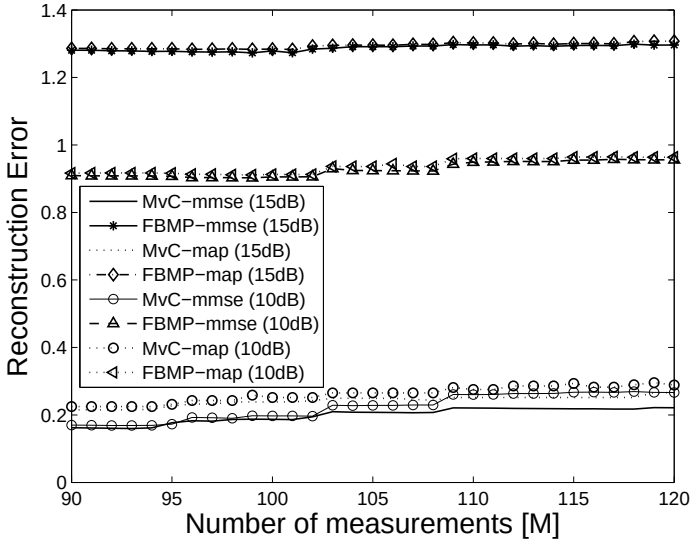
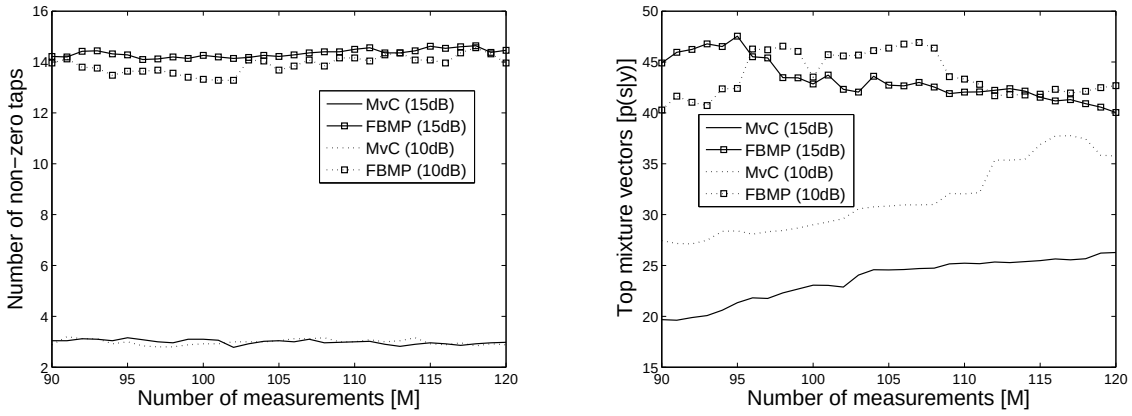


Figure 5.2: Average MMSE and MAP reconstruction errors for FBMP and MvC methods as a function of M (SNR=10, 15 dB).

Figure 5.3(a) shows the average number of non-zero taps contained in the most significant basis configurations for both methods. Our proposed method results in a decreased number of activated taps. On the other hand, Figure 5.3(b) shows the average number of vectors contained in the set $\mathcal{T}_{s,l_{max}}$ consisting of the most significant basis configurations (mixture vectors). Our method exploits a smaller number of such configurations, while for a smaller SNR both methods require more mixture vectors to capture the impulsive behavior. In order to make the comparison of the sparsity performance of the two methods more meaningful we define the following sparsity ratio

$$\text{SpR} = \frac{(\#\text{ non-zero taps}) \times (\#\text{ significant mixture vectors})}{M} \ . \tag{5.31}$$

The lower the value of SpR, for a fixed M, the sparser the solution of the corresponding method.



(a) Average number of non-zero taps for FBMP and MvC methods as a function of M (SNR=10, 15 dB).

(b) Average number of significant mixture vectors for FBMP and MvC methods as a function of M (SNR=10, 15 dB).

Figure 5.3: Sparsity performance of FBMP and MvC methods.

Figure 5.4 shows the SpR ratio for the two methods and for the two SNR values. As it can be seen, for the same number of CS measurements the SpR ratio of the proposed method is much smaller than the corresponding value of the FBMP approach, which means that our proposed method results in a sparser solution. The fact that the value of SpR corresponding to the proposed method is almost constant over the whole range of M may be due to the fact that we do not re-estimate the mixture parameters (μ_i, σ_i^2) during the reconstruction process. This may affect the sensitivity of the proposed algorithm.

As it was mentioned before, the degree of impulsiveness of the signal under consideration is controlled by the value of the characteristic exponent α of the sub-Gaussian $S\alpha S$ distribution. Although in our proposed model we consider the multivariate Cauchy case ($\alpha = 1$) we are interested in studying the reconstruction performance for other values of α as well, corresponding to different levels of impulsiveness. For this purpose, we curry out a second set of Monte-Carlo runs by generating simulated measurement vectors $\vec{g}$ according to Eq. (5.1), where the sparse vectors $\vec{w}$ are drawn from a sub-Gaussian distribution of length $N = 300$ that contain L spikes, whose locations are chosen at random. As before, the sparsity is set as a function of λ_1 and N, $L = \lceil \lambda_1 \cdot N \rceil$, with $\lambda_1 = 0.03$. The measurement noise is generated by drawing samples from a zero-mean sub-Gaussian distribution with underlying standard deviation σ_η. We fix the number of measurements to $M = 90$ and the $M \times N$ measurement matrix $\mathbf{\Phi}$ is constructed by first drawing i.i.d. samples from a standard Gaussian distribution and then by normalizing its columns to unit magnitude.

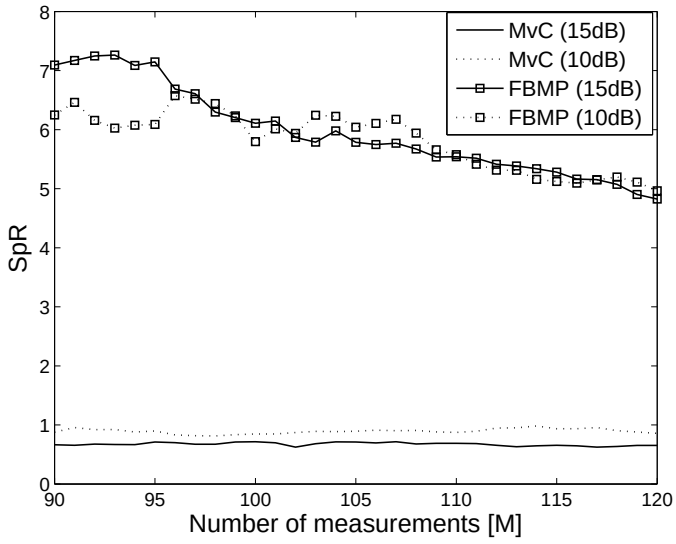


Figure 5.4: SpR ratio for FBMP and MvC methods as a function of M (SNR=10, 15 dB).

The reconstruction performance is tested by varying the characteristic exponent α in the interval $[1, 2]$ with a step size of 0.1 and for two SNR values, SNR=8, 10 dB. The underlying variance of the noise, σ_η^2, is set in accordance to the mixture variance σ_1^2 via Eq. (5.28). We also set the values of the mixing parameters μ_1 and σ_1^2 to 3 and 5, respectively. The process is repeated for 100 independent Monte-Carlo realizations and the results are given by averaging over the 100 runs.

Figure 5.5 on the following page shows the $\text{NMSE}_{\text{MMSE}}$ and NMSE_{MAP} reconstruction errors, as a function of α, corresponding to the FBMP and MvC methods, for the two SNR values. First, we observe that the performance of both methods is improved as the SNR increases. Besides, the MvC approach achieves a smaller reconstruction error in comparison to the FBMP approach for values of α close to 1, that is, when the actual distribution of the signal is heavy-tailed, while its performance is comparable to the performance of the FBMP as α tends to 2 (that is, to a Gaussian prior model). Thus, when the original signal is highly sparse the MvC approach should be preferred.

Figure 5.6 shows the SpR ratio as a function of α for the FBMP and MvC methods, for the two SNR values. As before, for the values of α which are in the vicinity of 1 the SpR ratio of the proposed method is much smaller than the SpR ratio of the FBMP approach, which means that our proposed method results in a much sparser solution when working in a highly impulsive environment. On the other hand, as the value of α approaches 2 (Gaussian statistics) the SpR ratio of both methods decreases, with the FBMP method, which is based on a Gaussian prior, resulting in a slightly sparser solution.

5.5 Conclusions and future work

In this chapter, we introduced a method for CS reconstruction of a highly impulsive vector in heavy-tailed noise, developed in a Bayesian framework. We employed a multivariate Cauchy (MvC) distribution as the prior model, and thus modeling directly the vector $\vec{w}$ with a heavy-tailed distribution that enforces its sparsity. The experimental results revealed an improved performance of the proposed approach when compared with the previous FBMP method, which was designed under a Gaussian assumption. In particular, we showed that the MvC-based implementation achieves a smaller reconstruction error than the FBMP approach when the

observed signal is truly sparse (that is, $\alpha \rightarrow 1$), while maintaining a quite low value of the SpR ratio, which is equivalent to an increased sparsity.

In the present work, we made the simplified assumption that the components of a mixture vector $\vec{\tau}$ are chosen from two distributions ("inactive", "active"). Besides, the parameters of these distributions are predetermined and kept fixed during the reconstruction process. As a future work, we are interested in modifying the proposed model so as to permit each mixture component to be chosen from a larger set of candidate mixture distributions and also in introducing a technique for re-estimating their corresponding parameters (μ_i, σ_i^2) during the reconstruction process.

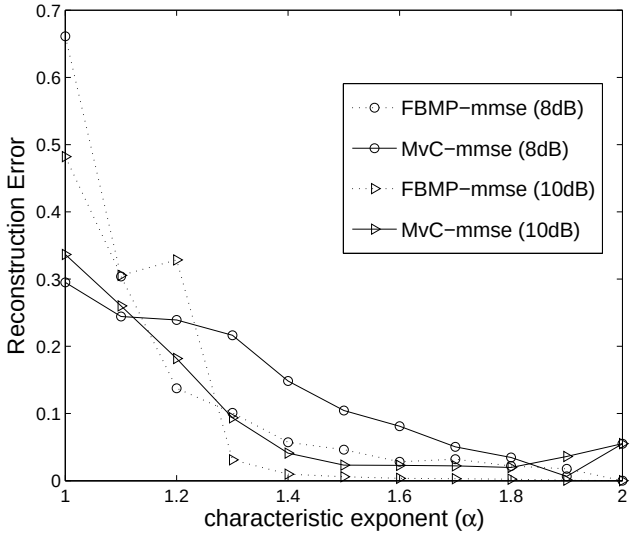


Figure 5.5: Average MMSE and MAP reconstruction errors for FBMP and MvC methods as a function of α (SNR=8, 10 dB).

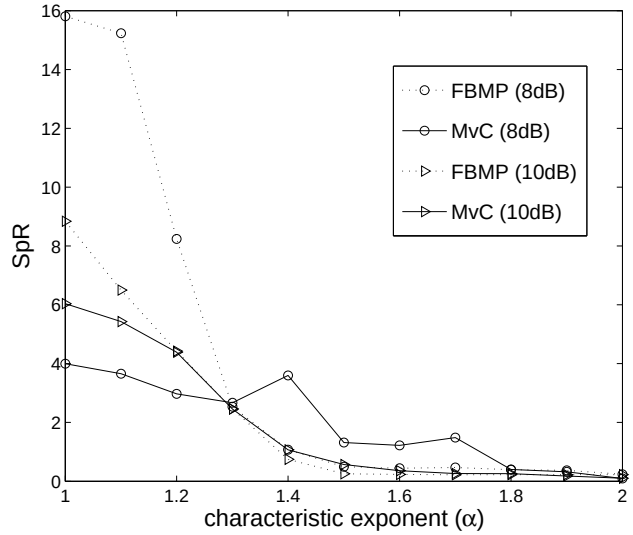


Figure 5.6: SpR ratio for FBMP and MvC methods as a function of α (SNR=8, 10 dB).

Chapter

6

Compressed Sensing using Symmetric Alpha-Stable Distributions

What most experimenters take for granted before they begin their experiments is infinitely more interesting than any results to which their experiments lead.

Norbert Wiener (1894-1964)
Mathematician

6.1 Introduction

Sensor networks gather an enormous amount of data over space and time to derive an estimate of a parameter or function from them. Several constraints, such as limited power, communication bandwidth, and storage capacity, motivate the need for a new paradigm for sensor data processing in order to extend the network's lifetime, while also obtaining accurate estimates. Besides, the Gaussian assumption has played a predominant role in signal processing being widely used as a signal and noise model. However, this assumption is very often unrealistic for a wide range of real-world data, which can be highly sparse in appropriate orthonormal bases.

In the present chapter, the inherent property of compressed sensing (CS) theory working simultaneously as a sensing and compression protocol using a small subset of random projections is exploited to reduce the total amount of data handled by each sensor. We propose a new iterative algorithm for reconstructing non-negative signals in highly impulsive environments, which can be modelled by members of the symmetric alpha-Stable distributions family. In addition, we develop a distributed implementation using duality theory and subgradient-based optimization. The subsequent performance evaluation reveals that our proposed method results in an increased reconstruction performance, while also achieving a high sparsity using much less basis functions when compared with state-of-the-art constrained optimization algorithms.

A major challenge in designing wireless sensor network (WSN) systems and algorithms is that transmitting data from a sensor to a central processing node may set a significant exhaustion of communication and energy resources. Such concerns may place undesirable limits on the amount of data collected and processed by sensor networks. Thus, it is natural to seek distributed algorithms for processing the data gathered by the nodes of a sensor network.

Several works [134, 6], have shown that many natural signals result in highly sparse representations when they are projected on localized orthonormal bases (e.g., wavelets, sinusoids). The traditional approach to compressing such sparse signals is to compute their transform coefficients and then store or transmit only a small number of large amplitude coefficients. However, this is an inherently wasteful process (in terms of both sampling rate and computational com-

plexity), since one gathers and processes the entire signal even though an exact representation is not required explicitly.

Compressed sensing (CS) [4] enables a potentially significant reduction in sampling and computation costs at a sensing system with limited capabilities. In particular, a signal having a sparse representation in a transform domain can be reconstructed from a small set of projections onto a second, measurement basis that is incoherent with the first one. Thus, CS provides a simple compression scheme with low computational complexity. Compressive wireless sensing (CWS) [72, 135] appears to be able to reduce the latency of data gathering in a single-hop network by delivering linear projections of sensor readings through synchronized amplitude-modulated analog transmissions or in a distributed fashion.

The majority of previous works on CS-based reconstruction of a sparse signal solve constrained optimization problems. Commonly used approaches are typically based on convex relaxation (Basis Pursuit [37]), non-convex (gradient based) local optimization (Re-weighted ℓ_1 minimization [136]) or greedy strategies ((Orthogonal) Matching Pursuit ((O)MP [137, 138]). The increased computation cost per iteration of those methods was addressed by introducing a general framework, the so-called Gradient Pursuits (GP) [36].

All these methods have been applied in scenarios where the underlying process generating the signal and/or the noise follows a Gaussian model. However, the normality assumption is violated in several distinct environments, such as in underwater acoustics [139], in sonar/radar [140], in telephony/satellite communications [141] and in finance [142], where the associate signals and/or noise take large-amplitude values much more frequently than what a Gaussian model implies.

For this purpose, we develop a new iterative greedy algorithm for CS reconstruction of sparse signals corrupted by additive heavy-tailed noise. In particular, the impulsive behavior is modelled using members from the family of *symmetric alpha-Stable* ($S\alpha S$) distributions, which are heavy-tailed and thus appropriate for representing highly impulsive phenomena. Then, this method is extended in an efficient distributed fashion, such as to be applicable in a wireless sensor network by reducing significantly the total amount of data handled by each sensor.

For convenience, in the next section we review briefly the statistical signal model employed by the proposed method, as well as some of the main properties of the family of $S\alpha S$distributions introduced in Chapter 3.

6.2 Statistical signal model

Let $\mathbf{\Psi}$ be a $N \times N$ matrix, whose columns correspond to the transform basis functions. Then, a given signal $\vec{f} \in \mathbb{R}^N$ can be represented as $\vec{f} = \mathbf{\Psi}\vec{w}$, where $\vec{w} \in \mathbb{R}^N$ is the weight vector. Obviously, $\vec{f}$ and $\vec{w}$ are equivalent representations of the original signal in the space and the transform domain, respectively. As mentioned above, for many real-world signals the majority of the components of $\vec{w}$ have negligible amplitude. In particular, $\vec{f}$ is L-sparse in basis $\mathbf{\Psi}$ if the corresponding weight vector $\vec{w}$ has exactly L non-zero components ($L \ll N$). In a real-world scenario $\vec{f}$ is not strictly L-sparse, but it is said to be *compressible* when the re-ordered components of $\vec{w}$ decay at a power-law.

Consider also an $M \times N$ ($M < N$) measurement matrix $\mathbf{\Phi}$, where the rows of $\mathbf{\Phi}$ are incoherent with the columns of $\mathbf{\Psi}$. For instance, let $\mathbf{\Phi}$ be a Hadamard matrix or a matrix containing independent and identically distributed (i.i.d.) Gaussian entries. Such matrices are incoherent with any fixed transform matrix $\mathbf{\Psi}$ with high probability (universality property) [4].

If $\vec{f}$ is compressible in $\mathbf{\Psi}$, then, it is possible to perform directly a compressed set of measurements $\vec{g}$, resulting in a simplified acquisition system. The original signal $\vec{f}$ is related to the CS measurements $\vec{g}$ through random projections, $\vec{g} = \mathbf{\Phi}\mathbf{\Psi}^T\vec{f} = \mathbf{\Phi}\vec{w}$, where $\mathbf{\Phi} = [\vec{\phi}_1, \ldots, \vec{\phi}_M]^T$

and $\vec{\phi}_m \in \mathbb{R}^N$ is a random vector with i.i.d. components. Thus, the reconstruction of $\vec{f}$ from $\vec{g}$ reduces to estimating the sparse weight vector $\vec{w}$.

Most of the recent literature on CS has concentrated on solving constrained optimization problems for signal reconstruction using a set of noisy CS measurements,

$$\vec{g} = \mathbf{\Phi}\vec{w} + \vec{\eta}\ , \tag{6.1}$$

where the unknown vector $\vec{w}$ and/or the noise $\vec{\eta}$ (with unknown variance σ_η^2) are modelled as Gaussian random variables. However, the Gaussian assumption is not efficient for a highly sparse coefficient vector $\vec{w}$. For this purpose, recent works incorporated several non-Gaussian (heavy-tailed) distributions (e.g., Student-t, Laplace, Cauchy, GSM) [50, 143, 144, 145] for modelling the prior belief that the vast majority of $\vec{w}$'s components have negligible amplitude.

In the present study, the prior belief that $\vec{w}$ is highly sparse is exploited by using a $S\alpha S$ distribution as its prior, which is heavy-tailed and thus, suitable for representing an impulsive behavior. Notice that in this case a high sparsity can be viewed as a high impulsiveness in the sense that only a very small number of $\vec{w}$'s components deviates from zero (all the components with a negligible amplitude can be considered to be "zero"). In the following discussion we consider that the noise $\vec{\eta}$ is also drawn from a $S\alpha S$ distribution. The use of this family is also motivated by the fact that the tails of a $S\alpha S$ distribution decay at an algebraic rate, which is in agreement with the rate of decay of the re-ordered components of the (compressible) vector $\vec{w}$.

In the following we revisit for convenience some of the fundamental properties of the family of univariate $S\alpha S$ distributions exploited in the design of the proposed CS method. As it was mentioned in Chapter 3, a $S\alpha S$ distribution is best defined by its characteristic function [112]:

$$\phi(t) = \exp(\imath\delta t - \gamma^\alpha |t|^\alpha), \tag{6.2}$$

where α is the *characteristic exponent*, taking values $0 < \alpha \leq 2$, δ ($-\infty < \delta < \infty$) is the *location parameter* and γ ($\gamma > 0$) is the *dispersion* of the distribution. The characteristic exponent is a shape parameter, which controls the "thickness" of the tails of the density function. The smaller the α, the heavier the tails of the $S\alpha S$ density function. The dispersion parameter determines the spread of the distribution around its location parameter, similar to the variance of the Gaussian. A $S\alpha S$ distribution is called *standard* if $\delta = 0$ and $\gamma = 1$. The notation $X \sim f_\alpha(\gamma, \delta)$ means that the random variable X follows a $S\alpha S$ distribution with parameters α, γ, δ.

In general, no closed-form expressions exist for most $S\alpha S$ density and distribution functions except for the Gaussian ($\alpha = 2$) and the Cauchy ($\alpha = 1$). As mentioned before, unlike the Gaussian density which has exponential tails, stable densities have tails following an algebraic rate of decay ($P(X > x) \sim Cx^{-\alpha}$, as $x \to \infty$, where C is a constant depending on the model parameters), hence $S\alpha S$ random variables with small α values are highly impulsive.

An important characteristic of $S\alpha S$ distributions is the non-existence of second-order moments. Instead, all moments of order p less than α do exist and are called the *Fractional Lower Order Moments* (FLOMs). In particular, the FLOMs of a $S\alpha S$ random variable $X \sim f_\alpha(\gamma, \delta = 0)$ are given by Eq. (3.6):

$$\mathbb{E}\{|X|^p\} = \big(C(p,\alpha)\cdot\gamma\big)^p\ , \quad 0 < p < \alpha, \tag{6.3}$$

where

$$\big(C(p,\alpha)\big)^p = \frac{2^{p+1}\Gamma\Big(\frac{p+1}{2}\Big)\Gamma\Big(-\frac{p}{\alpha}\Big)}{\alpha\sqrt{\pi}\,\Gamma\Big(-\frac{p}{2}\Big)} = \frac{\Gamma\big(1-\frac{p}{\alpha}\big)}{\cos\big(\frac{\pi}{2}p\big)\Gamma(1-p)}\ . \tag{6.4}$$

The $S\alpha S$ model parameters (α, γ) can be estimated using the consistent Maximum Likelihood (ML) method described by Nolan [113], which gives reliable estimates and provides the tightest possible confidence intervals. In addition, from Eq. (3.6) we obtain the following expression for the dispersion of X:

$$\gamma_X = \frac{\big(\mathbb{E}\{|X|^p\}\big)^{1/p}}{C(p,\alpha)} \ . \tag{6.5}$$

The *covariation norm* of $X \sim f_\alpha(\gamma, 0)$ with $\alpha > 1$, is defined by $\|X\|_\alpha = \gamma_X$, where γ_X is given from Eq. (6.5). This definition is extended to a quasi-norm for $\alpha < 1$, resulting in the following expressions:

$$\|X\|_\alpha = \begin{cases} \gamma_X & \text{, for } 0 < p < \alpha,\ 1 \le \alpha \le 2 \\ \gamma_X^\alpha & \text{, for } 0 < p < \alpha,\ 0 < \alpha < 1 \end{cases} \tag{6.6}$$

The concept of covariance plays a fundamental role in the second-order moment theory. However, covariances do not exist in the family of $S\alpha S$ random variables, due to their infinite variance. Instead, a quantity called *covariation*, which under certain constraints plays an analogous role for $S\alpha S$ random variables to the one played by covariance for Gaussian random variables, has been proposed. Let X and Y be jointly $S\alpha S$ random variables with $1 < \alpha \le 2$, zero location parameters and dispersions γ_X and γ_Y, respectively. Then, for all $1 < p < \alpha$ the covariation of X with Y was defined by Eq. (3.13) as follows:

$$[X, Y]_\alpha = \frac{E\{XY^{<p-1>}\}}{E\{|Y|^p\}} \|Y\|_\alpha^\alpha \ , \tag{6.7}$$

where for any real number z and $a \ge 0$ we use the notation

$$z^{<a>} = |z|^a sign(z) \ , \tag{6.8}$$

while for a real vector $\vec{z} \in \mathbb{R}^N$ and $a \ge 0$ we write

$$\vec{z}^{<a>} = [|z_1|^a sign(z_1), \ldots, |z_N|^a sign(z_N)] \ . \tag{6.9}$$

The covariation satisfies the following (pseudo-)linearity properties in the first and second argument, respectively (ref. Eqs. (3.14)-(3.15)): If X_1, X_2, Y are jointly $S\alpha S$ then

$$[aX_1 + bX_2, Y]_\alpha = a[X_1, Y]_\alpha + b[X_2, Y]_\alpha \tag{6.10}$$

$$[Y, aX_1 + bX_2]_\alpha = a^{<\alpha-1>}[Y, X_1]_\alpha + b^{<\alpha-1>}[Y, X_2]_\alpha \tag{6.11}$$

for any constants $a, b \in \mathbb{R}$.

If $X \sim f_\alpha(\gamma_X, 0)$ and $Y \sim f_\alpha(\gamma_Y, 0)$ are two independent $S\alpha S$ random variables then $cX \sim f_\alpha(|c|\gamma_X, 0)$ $(c \neq 0)$ and $X + Y \sim f_\alpha\big((\gamma_X^\alpha + \gamma_Y^\alpha)^{1/\alpha}, 0\big)$. Thus, for the noisy CS measurements $\vec{g} = \mathbf{\Phi}\vec{w} + \vec{\eta}$, if $\{w_i \sim f_\alpha(\gamma_i, 0)\}_{i=1}^N$ and $\{\eta_j \sim f_\alpha(\gamma_\eta, 0)\}_{j=1}^M$, then

$$g_j \sim f_\alpha\Big(\big[\sum_{i=1}^N (|\phi_{ji}|\gamma_i)^\alpha + \gamma_\eta^\alpha\big]^{1/\alpha}, 0\Big) \ , \ j = 1, \ldots, M \ , \tag{6.12}$$

where ϕ_{ji} is the element of $\mathbf{\Phi}$ in the j-th row and the i-th column, that is, each CS measurement also follows a $S\alpha S$ distribution with the same α.

6.2.1 A new $S\alpha S$ measurement matrix

The majority of previous CS methods employs a measurement matrix $\mathbf{\Phi}$ whose entries are i.i.d. draws of a zero-mean Gaussian distribution. In this section, we examine the performance of existing measurement matrices in the case of $S\alpha S$ underlying statistics and we proposed a new $S\alpha S$ measurement matrix. We carry out a set of 100 Monte-Carlo runs for each $\alpha \in [0.9 : 0.05 : 2]$, where in each run we generate distinct vectors $\vec{w} \in \mathbb{R}^{1000}$, $\vec{\eta} \in \mathbb{R}^{100}$ with i.i.d. $S\alpha S$ entries, where the dispersions are chosen randomly (uniformly) in the interval $[0.01, 2.5]$. We employ four measurement matrices: 1) $\mathbf{\Phi_1}$ whose entries are i.i.d zero-mean standard Gaussian samples ($S\alpha S$ with $\alpha = 2$) and normalized columns to unit ℓ_2-norm, 2) $\mathbf{\Phi_2}$ with entries being i.i.d standard $S\alpha S$ samples, 3) $\mathbf{\Phi_3}$ resulting by normalizing the columns of $\mathbf{\Phi_2}$ to unit ℓ_2-norm and 4) $\mathbf{\Phi_4}$ obtained by normalizing the columns of $\mathbf{\Phi_2}$ to unit covariation norm. Figure 6.1 shows that the value of α, resulting by averaging the α values estimated from $\vec{g}$ over all Monte-Carlo runs, approximates the true α when the matrix $\mathbf{\Phi_2}$ is employed, that is, the non-normalized matrix with entries drawn from a standard $S\alpha S$ distribution. Thus, the proposed CS method is developed by employing non-normalized measurement matrices with i.i.d. $S\alpha S$ entries as the most suitable ones. Most importantly, the choice of $\mathbf{\Phi_2}$ instead of a Gaussian matrix offers an additional degree of freedom (the value of α), resulting in a more efficient adaptation of the CS scheme to the sparsity of $\vec{w}$.

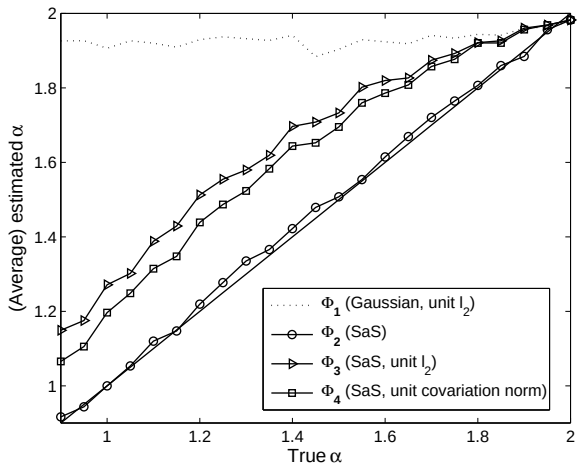


Figure 6.1: Test stability property for four kinds of measurement matrices $\mathbf{\Phi}$ (ref. in the text).

6.3 $S\alpha S$ minimum dispersion CS inversion

A CS-based inversion scheme aims at recovering the sparse vector $\vec{w}$ given the measurement matrix $\mathbf{\Phi}$ (in the subsequent analysis we use $\mathbf{\Phi} = \mathbf{\Phi}_2$) and the CS measurements $\vec{g}$. In the ideal case, we should look for a vector $\vec{w}$ with the smallest number of non-zero components, that is, with the smallest ℓ_0 norm. Although the problem of finding such a $\vec{w}$ is NP-hard, there exist several sub-optimal strategies which are used in practice. Most of them solve a constrained optimization problem by employing ℓ_2 or ℓ_1 norms. On the other hand, CS reconstruction methods were developed in recent works [146, 151] by employing ℓ_p norms with $p < 1$, with the goal of approximating the ideal ℓ_0 case.

Notice that for α-Stable data the Minimum Mean Squared Error (MMSE) criterion can be replaced by the *Minimum Dispersion* (MD) criterion in estimation problems since, unlike the

variance, the dispersion of an α-Stable random variable is finite and gives a good measure of the spread of estimation errors around zero. This provides a natural justification for the eligibility of ℓ_p norms with $p < 1$ in conjunction with a $S\alpha S$ prior model. From Eq. (6.5) we also observe that the MD criterion can be viewed as a *least ℓ_p-norm estimation error* criterion since the FLOM $\mathbb{E}\{|X|^p\}$ can be estimated as the ℓ_p norm of the vector X.

In the following, we develop an algorithm for approximating iteratively a sparse vector $\vec{w}$ with *non-negative* components. This is the case, for instance, of a signal with a sparse Fourier spectrum. In the subsequent derivations we will also use the following conventions: i) the set $\mathcal{I}^n$ contains n indices which indicate the n "active" columns of $\mathbf{\Phi}$ selected in the current iteration, ii) $\mathbf{\Phi}_{\mathcal{I}^n}$ denotes the sub-matrix of $\mathbf{\Phi}$ containing only those columns with indices in $\mathcal{I}^n$ (the same notation is also used for vectors, that is, $\vec{w}_{\mathcal{I}^n}$), iii) $\vec{w}_{\mathcal{I}^n}^{n-1}$ refers to an $card(\mathcal{I}^n)$-dimensional vector as calculated in iteration $n-1$, that is, the elements in $\vec{w}_{\mathcal{I}^n}^{n-1}$ which are not in $\vec{w}_{\mathcal{I}^{n-1}}^{n-1}$ are set to zero ($card(\cdot)$ denotes the cardinality of a set).

Similarly to the MP and OMP algorithms, the proposed approach minimizes an objective function depending on the norm of the approximation error. However, in contrast to MP and OMP (as well as to other greedy CS methods) that employ the squared ℓ_2 norm of the error, our $S\alpha S$-CS method optimizes the following cost function which is based on the ℓ_p norm ($p < 1$) due to the lack of second-order moments:

$$J_p(\vec{w}) = \sum_{i=1}^{N} |w_i|^p \ , \ \vec{w} \in \mathbb{R}^N \ , \ 0 \leq p \leq 1 \ . \tag{6.13}$$

In particular, we are interested in minimizing $J_p(\cdot)$ over $\vec{w}$ in terms of the estimation error $r(\vec{w}) = \vec{g} - \mathbf{\Phi}\vec{w}$ (for convenience we will also use the notation $\vec{r}$ instead of $r(\vec{w})$). This minimization will be implemented using a form of *directional updates*. Specifically, in the n-th iteration the estimate of the sparse vector is updated by calculating a direction $\vec{v}_{\mathcal{I}^n}^n$ and a step-size μ^n as follows,

$$\vec{w}_{\mathcal{I}^n}^n = \vec{w}_{\mathcal{I}^n}^{n-1} + \mu^n \vec{v}_{\mathcal{I}^n}^n \ . \tag{6.14}$$

A popular directional optimization approach is the conjugate gradient method that is guaranteed to solve quadratic optimization problems (e.g., MP/OMP). The conjugate gradient method uses directional updates that are $\mathbf{G}$-conjugate to the previously chosen directions, where $\mathbf{G} = \mathbf{\Phi}^T\mathbf{\Phi}$ is the Gram matrix. A set of vectors $\{\vec{v}_1, \ldots, \vec{v}_N\}$ is said to be $\mathbf{G}$-conjugate if $\vec{v}_i^T \mathbf{G} \vec{v}_j = 0$, $\forall\, i \neq j$. Following this approach the step-size μ^n is calculated by:

$$\mu^n = \frac{\vec{r}^{n^T} \mathbf{\Phi}_{\mathcal{I}^n} \vec{v}_{\mathcal{I}^n}^n}{\vec{v}_{\mathcal{I}^n}^{n^T} \mathbf{\Phi}_{\mathcal{I}^n}^T \mathbf{\Phi}_{\mathcal{I}^n} \vec{v}_{\mathcal{I}^n}^n} \tag{6.15}$$

and accordingly, the updated residual is given by:

$$\vec{r}^n = \vec{r}^{n-1} - \mu^n \mathbf{\Phi}_{\mathcal{I}^n} \vec{v}_{\mathcal{I}^n}^n \ . \tag{6.16}$$

Instead of calculating the exact conjugate direction, a more computationally efficient suboptimal direction is computed as a combination of the current gradient and the previous direction only :

$$\vec{v}_{\mathcal{I}^n}^n = [\vec{\nabla} J_p]_{\mathcal{I}^n} + b^n \vec{v}_{\mathcal{I}^n}^{n-1} \ , \tag{6.17}$$

where b^n is a step-size parameter and $[\vec{\nabla} J_p]$ is the negative gradient vector of the cost function

with respect to $\vec{w}$:

$$\begin{aligned}[\vec{\nabla} J_p] &= -\vec{\nabla}_{\vec{w}}\big(J_p(r(\vec{w}))\big) = -[\vec{\nabla}_{\vec{w}} r(\vec{w})][\vec{\nabla}_{\vec{w}}(J_p)(r(\vec{w}))] \\ &= \mathbf{\Phi}^T |p|\, diag\big(|r_1|^{p-2}, \ldots, |r_M|^{p-2}\big)\vec{r}\,, \end{aligned} \tag{6.18}$$

where $diag\big(|r_1|^{p-2}, \ldots, |r_M|^{p-2}\big)$ is the $M \times M$ diagonal matrix with elements the components of the residual vector $\vec{r} = \vec{g} - \mathbf{\Phi}\vec{w}$.

In order to calculate the step-size b^n we introduce a statistical pseudo-orthogonality condition between two $S\alpha S$ random variables. If X, Y are two jointly Gaussian random variable, they are considered to be orthogonal if their covariance is equal to zero. Since only the FLOMs are finite for α-Stable variables, then, if X, Y are two jointly $S\alpha S$ random variables we consider them to be orthogonal if they have zero "inner product", which we define as follows:

$$(X, Y) = \|Y\|_\alpha^{2-\alpha}\, [X, Y]_\alpha\ . \tag{6.19}$$

In the proposed $S\alpha S$-CS algorithm, the step-size b^n is computed by requiring the new direction to be "orthogonal" to the previous one, that is, $(\vec{v}^n_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}) = 0$, where:

$$\begin{aligned}(\vec{v}^n_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}) &= \|\vec{v}^{n-1}_{\mathcal{I}^n}\|_\alpha^{2-\alpha} [\vec{v}^n_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}]_\alpha \\ &\overset{(6.17)}{=} \|\vec{v}^{n-1}_{\mathcal{I}^n}\|_\alpha^{2-\alpha} \big[[\vec{\nabla} J_p]_{\mathcal{I}^n} + b^n \vec{v}^{n-1}_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}\big]_\alpha \\ &\overset{(3.14)}{=} \|\vec{v}^{n-1}_{\mathcal{I}^n}\|_\alpha^{2-\alpha} \Big(\big[[\vec{\nabla} J_p]_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}\big]_\alpha + b^n \big[\vec{v}^{n-1}_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}\big]_\alpha \Big) \end{aligned}$$

By equating the last expression with zero and noting that $\|\vec{v}^{n-1}_{\mathcal{I}^n}\|_\alpha \neq 0$ and $[\vec{v}^{n-1}_{\mathcal{I}^n}, \vec{v}^{n-1}_{\mathcal{I}^n}]_\alpha = \gamma^\alpha_{\vec{v}^{n-1}_{\mathcal{I}^n}}$ we calculate the step-size b^n as follows:

$$b^n = -\frac{\mathbb{E}\{[\vec{\nabla} J_p]_{\mathcal{I}^n} .*\ (\vec{v}^{n-1}_{\mathcal{I}^n})^{<p-1>}\}}{\mathbb{E}\{|\vec{v}^{n-1}_{\mathcal{I}^n}|^p\}}\ , \tag{6.20}$$

where the expectations are estimated by taking the mean values of the corresponding vectors and ".∗" denotes element-by-element multiplication between two vectors.

6.3.1 Basis selection rule

A task that affects significantly the performance of a CS reconstruction algorithm is the appropriate selection of the sparsest subset of basis vectors (columns of $\mathbf{\Phi}$) that best represents the data ($\vec{g}$). For instance, MP selects iteratively the column of $\mathbf{\Phi}$ resulting in the largest (in absolute magnitude) inner product with the current approximation error (residual) $\vec{r}^n$, which is equivalent to the largest reduction of $\vec{r}^n$. On the other hand, OMP selects the column of $\mathbf{\Phi}$ which, together with the already selected columns, yields the best signal approximation in the current iteration. In a recent work [152], the single-column selection rule was extended in a more efficient implementation which reduces the overall number of iterations by select *all columns* that satisfy a certain condition. For instance, in the n-th iteration one selects all columns whose inner product with $\vec{r}^n$ is above a threshold defined as a function of the norm of the residual.

In the proposed $S\alpha S$-CS algorithm, we select the optimal set of basis functions in the n-th iteration, $\mathcal{I}^n$, by introducing the following "distance measure" between two $S\alpha S$ random vectors $\vec{x}, \vec{y} \in \mathbb{R}^N$, based on FLOMs (with $0 < p < \alpha$):

$$d_\alpha(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|_\alpha = \begin{cases} \frac{\left(\mathbb{E}\{|\vec{x}-\vec{y}|^p\}\right)^{1/p}}{C(p,\alpha)} & \text{, for } 1 \leq \alpha \leq 2 \\ \frac{\left(\mathbb{E}\{|\vec{x}-\vec{y}|^p\}\right)^{\alpha/p}}{C(p,\alpha)} & \text{, for } 0 < \alpha < 1 \end{cases} \tag{6.21}$$

Algorithm 4 $S\alpha S$-CS estimation of a sparse vector $\vec{w}$

Input: $\mathbf{\Phi}$, $\vec{g}$, ξ, ϵ

Initialize: $\vec{r}^0 = \vec{g}$, $\vec{w}^0 = \vec{0}$, $\mathcal{I}^0 = \emptyset$

ML estimate of α from $\vec{g}$
Find optimal p (as function of α)
for $n = 1$; $n = n + 1$ until convergence criterion is met, **do**
  $[\vec{\nabla} J_p]^n = \mathbf{\Phi}^T |p| \, diag\big(|r_1^n|^{p-2}, \ldots, |r_M^n|^{p-2}\big) \vec{r}^n$
  $\mathcal{I}^n = \mathcal{I}^{n-1} \cup \{i \mid d_\alpha(\vec{\phi}_i, \vec{r}^n) \leq \xi \cdot \min_j (d_\alpha(\vec{\phi}_j, \vec{r}^n))\}$
  if $n = 1$ **then**
    $\vec{v}^1_{\mathcal{I}^1} = 1$
  end if
  if $n \neq 1$ **then**
    $b^n = -\mathbb{E}\{[\vec{\nabla} J_p]_{\mathcal{I}^n} .* (\vec{v}^{n-1}_{\mathcal{I}^n})^{<p-1>}\} / \mathbb{E}\{|\vec{v}^{n-1}_{\mathcal{I}^n}|^p\}$
    $\vec{v}^n_{\mathcal{I}^n} = [\vec{\nabla} J_p]^n_{\mathcal{I}^n} + b^n \vec{v}^{n-1}_{\mathcal{I}^n}$
  end if
  $\vec{q}^n = \mathbf{\Phi}_{\mathcal{I}^n} \vec{v}^n_{\mathcal{I}^n}$
  $\mu^n = (\vec{r}^{n^T} \vec{q}^n) / (\vec{q}^{n^T} \vec{q}^n)$
  $\vec{w}^n_{\mathcal{I}^n} = \vec{w}^{n-1}_{\mathcal{I}^n} + \mu^n \vec{v}^n_{\mathcal{I}^n}$
  $\vec{r}^n = \vec{r}^{n-1} - \mu^n \vec{q}^n$
end for

where $|\vec{x} - \vec{y}|^p = (|x_1 - y_1|^p, \ldots, |x_N - y_N|^p)$. In particular, an index i is added in $\mathcal{I}^n$ if the "distance" between the i-th basis vector, $\vec{\phi}_i$, and the current residual, $\vec{r}^n$, is below a certain threshold,

$$\mathcal{I}^n = \mathcal{I}^{n-1} \cup \{i \mid d_\alpha(\vec{\phi}_i, \vec{r}^n) \leq \xi \cdot \min_j (d_\alpha(\vec{\phi}_j, \vec{r}^n))\} \ , \tag{6.22}$$

with $\xi > 1$ (usually it suffices $\xi \approx (1 + \delta)$ with δ being close to zero). Notice that when $\xi = 1$ the basis selection rule reduces to the completely greedy approach, where a single optimal basis vector is selected in each iteration.

The algorithm terminates whether a predefined maximum number of iterations is reached, or when the relative decrease of the residual ℓ_p-norm falls bellow a threshold,

$$\frac{\|\vec{r}^n\|_p^p - \|\vec{r}^{n-1}\|_p^p}{\|\vec{r}^{n-1}\|_p^p} < \epsilon \ , \quad \epsilon \ll 1 \ .$$

Algorithm 4 summarizes the proposed $S\alpha S$-CS reconstruction technique.

6.3.2 Estimation of p parameter

Notice that most of the quantities involved in the proposed $S\alpha S$-CS algorithm depend on the specification of a parameter p, whose optimal value is a function of the characteristic exponent α. We compute the optimal p value as the one that *minimizes the standard deviation* of the FLOM-based covariation estimator, obtained from Eq. (3.13) by replacing the expectations with the arithmetic mean, for different values of $\alpha > 1$. For this purpose we studied the influence of p on the performance of the covariation estimator via Monte-Carlo runs using two $S\alpha S$ random variables X, Y of length $N = 2048$ (N is chosen to be a power of 2 for a faster ML estimation of the $S\alpha S$ parameters). We executed 1000 Monte-Carlo runs with $\alpha \in [0.9 : 0.05 : 2]$ (for computational reasons we selected to start with $\alpha = 0.9$ which is close to but less than 1) and for dispersions (γ_X, γ_Y) ranging in the interval $[0.01, 5]$ with a denser sampling in the sub-interval $[0.01, 2.5]$.

Table 3.1 shows the averaged optimal p values (in that table p is denoted by q) as a function of $\alpha \geq 1$, where the average is taken over the corresponding optimal p's for all dispersion pairs (γ_X, γ_Y) and Monte-Carlo runs. This table is then used as a lookup table in order to find the optimal p for every $0.9 \leq \alpha \leq 2$ by interpolating these values or by extrapolating them when the numerical ML estimation results in $\alpha < 0.9$.

6.4 Distributed $S\alpha S$-CS

There are cases where the information, namely, the measurement matrix $\mathbf{\Phi}$ and the sparse signal $\vec{w}$, may not be available in a single node. For instance, consider the case of a sensor network where each sensor has access only to a portion of $\mathbf{\Phi}$. In the following, we assume that the columns of $\mathbf{\Phi}$ are distributed across the nodes/sensors of a network. Working in this framework, it is well known that a famous basis selection method, namely, the Basis Pursuit (BP) can be reformulated as a distributed linear program [147]. However, the resulting approach requires a fully connected network in the sense that, at each iteration, every sensor must be able to communicate with all the remaining ones. Motivated by this fact, we present a method for solving the $S\alpha S$-CS problem in a distributed fashion, while requiring a less demanding network topology.

The problem under consideration is stated as follows: "given K nodes each one storing a subset of columns of $\mathbf{\Phi}$, find appropriate network topologies along with distributed algorithms for solving the following problem (P1)",

PRIMAL ($\mathbf{P1}$):
$$\begin{aligned} \min \quad & J_p(\vec{w}) \\ \text{s.t.} \quad & \vec{g} = \mathbf{\Phi}\vec{w} + \vec{\eta} \\ & -\vec{w} \leq 0 \ , \end{aligned}$$

where by $-\vec{w} \leq 0$ ($\Leftrightarrow \vec{w} \geq 0$) we mean that each component of $\vec{w}$ should be non-negative. Notice also that now the objective function $J_p(\cdot)$ is applied directly on the sparse vector $\vec{w}$ instead of the residual $\vec{r}$ as before.

The columns of $\mathbf{\Phi}$ are distributed among K nodes, such that the k-th node stores the k-th submatrix in the horizontal partition $\mathbf{\Phi} = [\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_k, \ldots, \mathbf{\Phi}_K]$, where $\mathbf{\Phi}_k \in \mathbb{R}^{M \times n_k}$, and $n_1 + \cdots + n_K = N$. A corresponding partition also holds for the sparse vector, $\vec{w} = [\vec{w}_1, \ldots, \vec{w}_k, \ldots, \vec{w}_K]$, where $\vec{w}_k \in \mathrm{R}^{n_k}$. The proposed method is based on the *Duality Theory* [148]. Under the appropriate conditions, such as, separability of the objective function and the constraints, dual problems can be confronted by distributed methods. Hereafter, we assume that $\mathbf{\Phi}$ has full rank in order to ensure the feasibility of (P1) with high probability.

Unless the opposite is clearly stated, we assume throughout this section the following:
Feasibility & Boundedness: There exists at least one feasible solution for the primal problem (P1) and the objective function is bounded from below, that is, $-\infty < J_p^* < \infty$, where J_p^* denotes the optimal value of $J_p(\vec{w})$.

In order to enforce the unconstrained character of the corresponding dual problem, and therefore a direct applicability in a distributed setting (since the objective function is already separable), we introduce a redundant constraint. In particular, let $U > 0$ be an upper bound of the ℓ_∞ norm of any solution of (P1). Then, the bounded version of (P1) is given by:

Bounded PRIMAL ($\mathbf{P2}$):
$$\begin{aligned} \min \quad & J_p(\vec{w}) \\ \text{s.t.} \quad & \vec{g} = \mathbf{\Phi}\vec{w} + \vec{\eta} \\ & -\vec{w} \leq 0 \\ & \|\vec{w}\|_\infty^p \leq U \ . \end{aligned}$$

We also note that such an upper bound U can be easily found. Starting from the noiseless case, given any $\vec{w}$ such that $\vec{g} = \mathbf{\Phi}\vec{w}$ (e.g., $\vec{w} = \mathbf{\Phi}^{\dagger}\vec{g}$, where $\mathbf{\Phi}^{\dagger}$ denotes the pseudoinverse), it suffices to select $U > N\|\vec{w}\|_{\infty}^{p}$. Indeed, if $\vec{w}^{*}$ is a solution of (P1) the following inequalities hold:

$$\|\vec{w}^{*}\|_{\infty}^{p} = (\max_{1\leq i\leq N}\{|w_i^{*}|\})^{p} \leq J_p(\vec{w}^{*}) \leq J_p(\vec{w}) \leq N\|\vec{w}\|_{\infty}^{p} < U \ .$$

Generalizing in the noisy case, the following inequalities determine a rule for the selection of U:

$$\begin{aligned} \|\vec{w}^{*}\|_{\infty}^{p} &\leq J_p(\vec{w}^{*}) \leq J_p\big(\vec{w}\big) = J_p\big(\mathbf{\Phi}^{\dagger}(\vec{g}-\vec{\eta})\big) \\ &\leq N\|\mathbf{\Phi}^{\dagger}(\vec{g}-\vec{\eta})\|_{\infty}^{p} \leq N\|\mathbf{\Phi}^{\dagger}\|_{max}\|\mathbf{1}(\vec{g}-\vec{\eta})\|_{\infty}^{p} \\ &< (N\|\mathbf{\Phi}^{\dagger}\|_{max}R^{p}) = U \ , \end{aligned} \tag{6.23}$$

where $\|\mathbf{\Phi}^{\dagger}\|_{max} = \max\big\{\big|[\mathbf{\Phi}^{\dagger}]_{nm}\big|\big\}_{1\leq n\leq N, 1\leq m\leq M}$, $\mathbf{1} \in \mathbb{R}^{N\times M}$ is the matrix with all of its entries being equal to one and R is a positive constant greater than the maximum amplitude component of $\vec{g}-\vec{\eta}$. Since they are $S\alpha S$ random vectors this maximum is unknown in advance, but it suffices to select an R that satisfies this requirement with "high-probability". In a specific signal processing application, there is usually some prior knowledge about the signal content so that we can achieve an appropriate choice for R by assigning a relatively large value to it in comparison with the entries of the (known) measurement vector $\vec{g}$ and the expected noise amplitude.

6.4.1 Dualization and distributed solution of the primal problem

We consider the *dual function* $\mathcal{L}(\cdot)$ defined for $\vec{\lambda} \in \mathbb{R}^{M}$ as follows:

$$\mathcal{L}(\vec{\lambda}) = \inf_{\vec{w}\in\mathbb{R}_{+}^{N}} \mathcal{L}(\vec{w}, \vec{\lambda}) \ , \tag{6.24}$$

where $\mathcal{L}(\vec{w}, \vec{\lambda})$ is the *Lagrangian function* and $\vec{\lambda}$ is the vector of *Lagrange multipliers*. The *dual problem* is defined by:

DUAL (**D1**):
$$\begin{aligned} \max \quad & \mathcal{L}(\vec{\lambda}) \\ \text{s.t.} \quad & \vec{\lambda} \geq 0 \ . \end{aligned}$$

Following the standard dualization approach on all constraints except for the redundant ones and exploiting the separability of the objective function, as well as the partition of $\mathbf{\Phi}$ and $\vec{w}$, the Lagrangian function is expressed as follows:

$$\begin{aligned} \mathcal{L}(\vec{w}, \vec{\lambda}) &= J_p(\vec{w}) + \vec{\lambda}^{T}(\vec{g} - \mathbf{\Phi}\vec{w}) \\ &= \vec{\lambda}^{T}\vec{g} + \sum\nolimits_{k=1}^{K}\big(J_p(\vec{w}_k) - \vec{\lambda}^{T}\mathbf{\Phi}_k\vec{w}_k\big) \ . \end{aligned} \tag{6.25}$$

Notice that although the noise component is not explicitly employed in the above expression, however its presence will always result in an approximation $\widehat{\vec{w}}^{*}$ of the optimal vector $\vec{w}^{*}$. From the above, the dual function given by:

$$\begin{aligned} \mathcal{L}(\vec{\lambda}) &= \vec{\lambda}^{T}\vec{g} + \sum\nolimits_{k=1}^{K}\mathcal{L}_k(\vec{\lambda}) \ , \quad \text{where} \\ \mathcal{L}_k(\vec{\lambda}) &= \inf\Big\{\big(J_p(\vec{w}_k) - \vec{\lambda}^{T}\mathbf{\Phi}_k\vec{w}_k\big) \ : \ \|\vec{w}_k\|_{\infty}^{p} \leq U\Big\}. \end{aligned} \tag{6.26}$$

Because of the lack of second-order statistics we are interested in developing a distributed $S\alpha S$-CS algorithm based on FLOMs. The standard Lagrangian function of Eq. (6.25) employs

the usual inner product, which can be viewed as a measure of variance between the associated vectors. For this purpose, we propose the use of a Lagrangian function that exploits (fractional lower-order) covariations instead of variances and thus, it adapts to our $S\alpha S$ framework. The proposed Lagrangian is given by the following equation:

$$\begin{aligned}\mathcal{L}^{\mathcal{S}}(\vec{w},\vec{\lambda}) &= J_p(\vec{w}) + (\vec{\lambda}, \vec{g} - \mathbf{\Phi}\vec{w}) \\ &\overset{(6.19)}{=} J_p(\vec{w}) + \underbrace{\|\vec{g} - \mathbf{\Phi}\vec{w}\|_{\alpha}^{2-\alpha}[\vec{\lambda}, \vec{g} - \mathbf{\Phi}\vec{w}]_{\alpha}}_{\text{Ⓢ}} \,. \end{aligned} \tag{6.27}$$

For convenience we will restrict ourselves to the case $1 \leq \alpha \leq 2$. By noting that $\|\vec{g} - \mathbf{\Phi}\vec{w}\|_{\alpha}^{2-\alpha} = \|\vec{\eta}\|_{\alpha}^{2-\alpha} = \gamma_{\eta}^{2-\alpha}$ (from Eq. (6.6)) and using the pseudo-linearity property of Eq. (3.15), the second term of Eq. (6.27) takes the following form:

$$\begin{aligned}\text{Ⓢ} &= \gamma_{\eta}^{2-\alpha}\Big([\vec{\lambda}, \vec{g}]_{\alpha} + (-1)^{<\alpha-1>}[\vec{\lambda}, \mathbf{\Phi}\vec{w}]_{\alpha}\Big) \\ &\overset{(3.13)}{=} \gamma_{\eta}^{2-\alpha}\left(\frac{\mathbb{E}\{\vec{\lambda} .* \vec{g}^{<p-1>}\}}{\mathbb{E}\{|\vec{g}|^p\}}\gamma_g^{\alpha} - \frac{\mathbb{E}\{\vec{\lambda} .* \left(\mathbf{\Phi}\vec{w}\right)^{<p-1>}\}}{\mathbb{E}\{|\mathbf{\Phi}\vec{w}|^p\}}\gamma_{\mathbf{\Phi}\vec{w}}^{\alpha}\right) \\ &\overset{(3.6)}{=} \gamma_{\eta}^{2-\alpha}\left(\frac{\mathbb{E}\{\vec{\lambda} .* \vec{g}^{<p-1>}\}}{C(p,\alpha)^p\,\gamma_g^{p-\alpha}} - \frac{\mathbb{E}\{\vec{\lambda} .* \left(\mathbf{\Phi}\vec{w}\right)^{<p-1>}\}}{C(p,\alpha)^p\,\gamma_{\mathbf{\Phi}\vec{w}}^{p-\alpha}}\right) \\ &= \gamma_{\eta}^{2-\alpha}\left(\frac{\mathbb{E}\{\vec{\lambda} .* \vec{g}^{<p-1>}\}}{C(p,\alpha)^p\,\gamma_g^{p-\alpha}} - \frac{\mathbb{E}\{\vec{\lambda} .* \big(\sum_{k=1}^{K}\mathbf{\Phi}_k\vec{w}_k\big)^{<p-1>}\}}{C(p,\alpha)^p\,\gamma_{\mathbf{\Phi}\vec{w}}^{p-\alpha}}\right) \,, \end{aligned} \tag{6.28}$$

where we note again that ".*" denotes element-by-element multiplication between two vectors. In order to avoid numerical instability caused by the estimation of dispersions γ_g and $\gamma_{\mathbf{\Phi}\vec{w}}$, we will consider scenarios where the signal power is greater than the noise power (analogous to a relatively high SNR assumption). In this case $\gamma_g \simeq \gamma_{\mathbf{\Phi}\vec{w}}$. In addition, by substituting the expectations with the corresponding arithmetic means (expressed as inner products), Eq. (6.28) takes the following form:

$$\text{Ⓢ} = \frac{\gamma_{\eta}^{2-\alpha}\,\gamma_g^{\alpha-p}}{MC(p,\alpha)^p}\left[\vec{\lambda}^T\Big(\vec{g}^{<p-1>} - \big(\sum_{k=1}^{K}\mathbf{\Phi}_k\vec{w}_k\big)^{<p-1>}\Big)\right] \,. \tag{6.29}$$

Since γ_{η} is unknown and also along with γ_g they act as positive scaling factors and thus, they do not affect the minimization operator, the final expression of the proposed Lagrangian function is given by:

$$\mathcal{L}^{\mathcal{S}}(\vec{w},\vec{\lambda}) = \sum_{k=1}^{K} J_p(\vec{w}_k) + \vec{\lambda}^T\Big(\vec{g}^{<p-1>} - \big(\sum_{k=1}^{K}\mathbf{\Phi}_k\vec{w}_k\big)^{<p-1>}\Big) \,, \tag{6.30}$$

which is in a separable form and thus, amenable to a distributed implementation. In particular, we have to solve the dual problem (D1) using the following dual function

$$\mathcal{L}^{\mathcal{S}}(\vec{\lambda}) = \inf_{\substack{\vec{w}\in\mathbb{R}_{+}^{N} \\ \|\vec{w}\|_{\infty}^{p}\leq U}} \mathcal{L}^{\mathcal{S}}(\vec{w},\vec{\lambda}) \,. \tag{6.31}$$

We do this by employing the method of subgradients [148], that is,

$$\vec{\lambda}^{i+1} = [\vec{\lambda}^i + s^i\vec{d}(\vec{\lambda}^i)]^{+} \,, \tag{6.32}$$

where $\vec{\lambda}^i$ is the estimated dual variable in the i-th iteration, $s^i > 0$ is a step-size parameter, $\vec{d}(\vec{\lambda}^i)$ is a supergradient[1] of the dual function $\mathcal{L}^{\mathcal{S}}(\cdot)$ and $[\cdot]^+$ denotes the projection of a vector on the non-negative halfplane (due to the constraint of (D1)).

According to the subgradient method, for a sufficiently small step-size s^i the distance of the current iterate, $\vec{\lambda}^{i+1}$, to the optimal solution is reduced. In practice, the convergence of the subgradient method is ensured using the following step-size:

$$s^i = \frac{c^i \left(\widehat{\mathcal{L}^{\mathcal{S}}}(\vec{\lambda}^i) - \mathcal{L}^{\mathcal{S}}(\vec{\lambda}^i)\right)}{\|\vec{d}(\vec{\lambda}^i)\|_2} \ , \tag{6.33}$$

where $\widehat{\mathcal{L}^{\mathcal{S}}}$ is an approximation to the (unknown) optimal dual solution, which can be estimated using the best current dual value

$$\widehat{\mathcal{L}^{\mathcal{S}}}(\vec{\lambda}^i) = \max_{0 \le i' \le i} \mathcal{L}^{\mathcal{S}}(\vec{\lambda}^{i'}) \ . \tag{6.34}$$

In Eq. (6.33), c^i is a number chosen such that it guarantees a diminishing step-size. This can be achieved by setting

$$c^i = \frac{1+\beta}{i+\beta} \ , \tag{6.35}$$

where β is a fixed positive integer.

Turning back into Eq. (6.32), for a given $\vec{\lambda}$ a supergradient $\vec{d}(\vec{\lambda})$ can be obtained by differentiating Eq. (6.30) as follows:

$$\vec{d}(\vec{\lambda}) = \vec{g}^{<p-1>} - \Big(\sum_{k=1}^{K} \mathbf{\Phi}_k \vec{w}_k^*(\vec{\lambda})\Big)^{<p-1>} \ , \tag{6.36}$$

where $\vec{w}_k^*(\vec{\lambda})$ is chosen such that it maximizes $\mathcal{L}^{\mathcal{S}}(\vec{\lambda})$ obtained by substituting Eq. (6.30) in Eq. (6.31). We select the $\{\vec{w}_k^*(\vec{\lambda})\}_{k=1}^K$ by employing a heuristic approach as follows: first, notice that in the current i-th iteration the term $\vec{\lambda}^{i^T} \vec{g}^{<p-1>}$ can be considered as a constant and thus, it suffices to find $\vec{w}_k^*(\vec{\lambda}^i)$ $(1 \le k \le K)$ such that the vector $\vec{w}^*(\vec{\lambda}^i) = [\vec{w}_1^*(\vec{\lambda}^i), \ldots, \vec{w}_K^*(\vec{\lambda}^i)]$ satisfies the expression

$$\vec{w}^*(\vec{\lambda}^i) = \arg \inf_{\substack{\vec{w} \in \mathbb{R}_+^N \\ \|\vec{w}\|_\infty^p \le U}} \Big(\sum_{k=1}^{K} J_p(\vec{w}_k) - \vec{\lambda}^{i^T} (\sum_{k=1}^{K} \mathbf{\Phi}_k \vec{w}_k)^{<p-1>}\Big) \ . \tag{6.37}$$

The following relations hold under the consideration $\vec{a} \le \vec{b} \Leftrightarrow a_i \le b_i \, , \forall \, i$ (the same holds for "$\ge$"):

$$\begin{aligned} \Big(\sum_{k=1}^{K} \mathbf{\Phi}_k \vec{w}_k\Big)^{<p-1>} \quad &= \quad \big[|v_1|^{p-1} sign(v_1), \ldots, |v_M|^{p-1} sign(v_M)\big]^T \\ &\underset{sign(\cdot) \le 1}{\le} \quad \big[|v_1|^{p-1}, \ldots, |v_M|^{p-1}\big]^T \end{aligned} \tag{6.38}$$

[1] The vector $\vec{h}$ is a supergradient (resp. subgradient) of a concave (resp. convex) function f at the point $\vec{x}$ if $\forall \, \vec{y}$, $f(\vec{y}) \le f(\vec{x}) + \vec{h}^T(\vec{y} - \vec{x})$ (resp. $f(\vec{y}) \ge f(\vec{x}) + \vec{h}^T(\vec{y} - \vec{x})$).

where

$$v_m = \sum_{k=1}^{K} \sum_{j=1}^{n_k} [\mathbf{\Phi}_k]_{mj} w_{kj} \ , \tag{6.39}$$

with $[\mathbf{\Phi}_k]_{mj}$ denoting the (mj)-th element of the submatrix $\mathbf{\Phi}_k$ and w_{kj} being the j-th component of $\vec{w}_k$. Taking the inner products of both sides of Eq. (6.38) with $\vec{\lambda}^i$ under the dual constraint $\vec{\lambda}^i \geq 0$ results in the following relations

$$\begin{aligned} \vec{\lambda}^{iT} \Big(\sum_{k=1}^{K} \mathbf{\Phi}_k \vec{w}_k \Big)^{<p-1>} &\leq \vec{\lambda}^{iT} \left[|v_1|^{p-1}, \ldots, |v_M|^{p-1} \right]^T \\ = \sum_{m=1}^{M} \lambda_m^i |v_m|^{p-1} &= \sum_{m=1}^{M} |(\lambda_m^i)^{\frac{1}{p-1}} v_m|^{p-1} \ . \end{aligned} \tag{6.40}$$

From Eqs. (6.39)-(6.40) we can see that the m-th component of the current Lagrange multiplier λ_m^i, risen to the power of $1/(p-1)$, multiplies the m-th row of each submatrix $\mathbf{\Phi}_k$. We are interested in finding a $\vec{w}^*(\vec{\lambda}^i)$ that minimizes Eq. (6.37) under the constraint $\|\vec{w}\|_\infty^p \leq U$. By combining with the inequality in Eq. (6.40), whose right-hand side consists of non-negative terms, we suggest that instead of finding a $\vec{w}^*(\vec{\lambda}^i)$ satisfying Eq. (6.37) we relax this requirement by searching for a $\vec{w}^*(\vec{\lambda}^i)$ such that:

$$\vec{w}^*(\vec{\lambda}^i) = \arg \inf_{\substack{\vec{w} \in \mathbb{R}_+^N \\ \|\vec{w}\|_\infty^p \leq U}} \Big(\sum_{k=1}^{K} J_p(\vec{w}_k) - \sum_{m=1}^{M} |(\lambda_m^i)^{\frac{1}{p-1}} v_m|^{p-1} \Big) \ , \tag{6.41}$$

where $\{v_m\}_{m=1}^M$ (Eq. (6.39)) depend explicitly on $\{\vec{w}_k^*(\vec{\lambda}^i)\}_{k=1}^K$ and the relaxation refers to the fact that the estimated $\vec{w}^*$ does not achieve exactly the infimum of Eq. (6.37) but a lower value, with our goal being to make this difference as small as possible. This relaxation has the advantage that we estimate $\vec{w}^*$ without the ambiguity of the $sign(\cdot)$ function.

Since both terms in the parentheses of Eq. (6.41) are non-negative the infimum of their difference will be equal to zero. Notice also that the second term implies that the parts of the partition of $\mathbf{\Phi}$ and $\vec{w}$ corresponding to the k-th sensor are distributed over M (additive) terms in a row-wise way (Eq. (6.39)). Thus, in order to enforce this contribution of the k-th sensor to be close to its associated objective function value, $J_p(\vec{w}_k)$, we keep only these components of $\vec{w}_k$ for which the sum of their coefficients over those M terms is non-negative, that is, for the k-th sensor the set of indices $\mathcal{T}_k^i$ corresponding to the active components in the i-th iteration is given by,

$$\mathcal{T}_k^i = \left\{ j \ : \ \sum_{m=1}^{M} (\lambda_m^i)^{\frac{1}{p-1}} [\mathbf{\Phi}_k]_{mj} \geq 0 \right\} \ , \ 1 \leq j \leq n_k \ . \tag{6.42}$$

Each sensor computes individually its set $\mathcal{T}_k^i$, which is then transmitted to the central node (Fusion Center). There, the single set of the current active components, $\mathcal{T}^i$, is obtained as the union of the K sets,

$$\mathcal{T}^i = \bigcup_{k=1}^{K} \mathcal{T}_k^i \ . \tag{6.43}$$

Finally, the current "optimal" vector $\vec{w}^*(\vec{\lambda}^i)$ is formed as follows:

$$[\vec{w}^*(\vec{\lambda}^i)]_n = \begin{cases} 0 & , \text{if } n \notin \mathcal{T}^i \\ U^{\frac{1}{p}} & , \text{if } n \in \mathcal{T}^i \ . \end{cases} \tag{6.44}$$

The above discussion indicates a natural network topology for the distributed implementation of the proposed subgradient method, as shown in Figure 6.2.

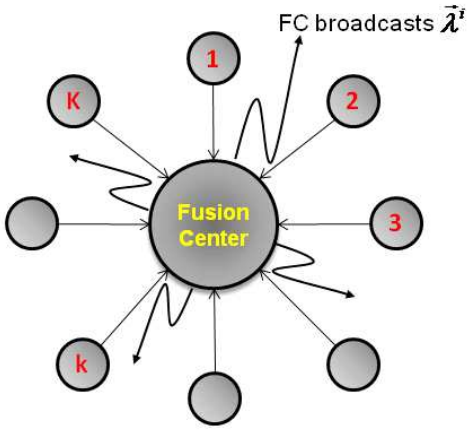


Figure 6.2: Network topology for implementing the distributed $S\alpha S$-CS algorithm.

Through the Karush-Kuhn-Tucker (KKT) conditions and the relaxation represented by Eq. (6.41), it can be seen that for any optimal solution $\vec{\lambda}^*$ of (D1), using $\mathcal{L}^{\mathcal{S}}(\cdot)$ instead of $\mathcal{L}(\cdot)$, we have that

$$supp\,(\vec{w}^*) \subset \mathcal{T}^* \;, \tag{6.45}$$

where $supp\,(\vec{w}^*) = \{n \,:\, [\vec{w}^*]_n \neq 0\}$ is the support of the optimal sparse vector satisfying (P2) and $\mathcal{T}^*$ is the final set of active components after the algorithm has converged. In other words, once the central node computes $\vec{\lambda}^*$ then, it obtains an over-estimate of the support of a solution of (P2) and thus, of (P1) since the two problems are equivalent. This means that at this point the central node could solve a problem (P1) of reduced dimensionality by removing the columns of $\mathbf{\Phi}$ whose indices are not included in $\mathcal{T}^*$ (and consequently setting $[\vec{w}]_{n'} = 0$ for $n' \notin \mathcal{T}^*$).

In particular, the central node estimates the sparse vector $\vec{w}$ satisfying the observation model

$$\vec{g} = \mathbf{\Phi}_{\mathcal{T}^*}\vec{w}_{\mathcal{T}^*} + \vec{\eta}\;, \tag{6.46}$$

using the $S\alpha S$-CS algorithm described in Section 6.3. Since the solution is expected to be highly sparse, this approach could also yield a significant reduction of the computational cost via the reduction in the problem dimensionality. The proposed distributed $S\alpha S$-CS strategy relies on the knowledge of $\vec{\lambda}^*$. In practice, the subgradient method is stopped at the central node after a maximum finite number of iterations is reached or when the relative error of the estimated dual variable falls below a predefined tolerance ε,

$$\|\vec{\lambda}^{i+1} - \vec{\lambda}^i\|_2 < \varepsilon \cdot \|\vec{\lambda}^{i+1} - \vec{\lambda}^0\|_2 \;. \tag{6.47}$$

As a result, the distributed algorithm converges to a suboptimal $\widehat{\vec{\lambda}}^*$ and consequently to a suboptimal set $\widehat{\mathcal{T}}^*$.

Notice that $\mathbf{\Phi}$ is distributed over the K nodes. However, $\mathbf{\Phi}_{\widehat{\mathcal{T}}^*}$ is required to the central node to estimate the sparse vector. This is carried out as follows: after stopping the subgradient method, the central node sends $\widehat{\vec{\lambda}}^*$ to the K nodes, which compute their corresponding fragments of $\widehat{\mathcal{T}}^*$ in a parallel way and transmit them back to the central node.

Concluding this section, we observe that the proposed distributed $S\alpha S$-CS algorithm consists of two stages: in the first stage the central node (FC) implements the subgradient method of

Algorithm 5 Distributed $S\alpha S$-CS estimation of a sparse vector $\vec{w}$

Input: Partitioned matrix $\{\mathbf{\Phi}_k\}_{k=1}^K$, $\vec{g}$ (available to FC), U, ξ, ϵ, β, MaxIter or ε (to terminate subgradient method)
Initialize: $\vec{\lambda}^0 = [1, \ldots, 1] \in \mathbb{R}^M$, $\vec{w}^0 = [0, \ldots, 0] \in \mathbb{R}^N$
ML estimate of α from $\vec{g}$
Find optimal p (as function of α)
for $i = 1$; $i = i + 1$ until convergence criterion is met, **do**
  Broadcast $\vec{\lambda}^i$ to the K sensors
  for each sensor $k = 1, \ldots, K$ **do**
    Estimate $\vec{w}_k^*(\vec{\lambda}^i)$ (from Eqs. (6.42), (6.44))
    Return $\mathbf{\Phi}_k \vec{w}_k^*(\vec{\lambda}^i)$ to the central node
  end for
  Compute $\vec{d}(\vec{\lambda}^i)$ (from Eq. (6.36)) at the central node
  Update Lagrange multiplier $\vec{\lambda}^{i+1}$ (using Eqs. (6.32)-(6.35))
end for
Broadcast $\widehat{\vec{\lambda}}^* = \vec{\lambda}^{i+1}$ to the K sensors
for each sensor $k = 1, \ldots, K$ **do**
  Compute $\mathcal{T}_k^*$ (from Eq. (6.42)) using $\widehat{\vec{\lambda}}^*$
  Return $\mathcal{T}_k^*$ and $[\mathbf{\Phi}_k]_{\mathcal{T}_k^*}$ to the central node
end for
At the central node form $\widehat{\mathcal{T}}^*$ (from Eq. (6.43)) and $\mathbf{\Phi}_{\widehat{\mathcal{T}}^*}$
Run reduced dimensionality Algorithm 4 with inputs $\mathbf{\Phi}_{\widehat{\mathcal{T}}^*}$, $\vec{g}$, ξ, ϵ
Output final sparse solution $\vec{w}^*$ at the central node

Eq. (6.32) to estimate and update $\widehat{\vec{\lambda}}^i$ by exchanging information with each sensor of the network in a parallel fashion. In the second stage, the final $\widehat{\vec{\lambda}}^*$ is broadcasted by the central node to the K sensors, which estimate efficiently their corresponding set of activated components $\mathcal{T}_k^*$ from Eq. (6.42) via simple additions and multiplications. Then, the central node forms $\widehat{\mathcal{T}}^*$ and $\mathbf{\Phi}_{\widehat{\mathcal{T}}^*}$ and solves a reduced dimensionality problem using the iterative $S\alpha S$-CS Algorithm 4. Algorithm 5 summarizes the proposed distributed $S\alpha S$-CS reconstruction method.

6.5 Performance evaluation

In this section, we evaluate the performance of the proposed $S\alpha S$-CS algorithm and its distributed version, by comparing it with state-of-the-art norm-based CS reconstruction methods. Of course there are quite many CS methods in the recent literature with which we could compare, however, the scope of this study is to highlight the advantages of the $S\alpha S$ model in developing CS reconstruction algorithms and to exhibit its superiority when compared with some of the most recent norm-based iterative CS methods. In particular, the following methods are used for comparisons, which employ ℓ_2 and ℓ_p ($p \leq 1$) norms: 1) Stagewise Orthogonal Matching Pursuit (StOMP) [39], 2) LASSO with a non-negativity constrained (nnLasso) [23], 3) ℓ_1-norm minimization using the primal-dual interior point method (L1EQ-PD), 4) Smoothed ℓ_0 (SL0) [150], 5) Affine Scaling Transformation with varying p diversity measure (AST) [151], 6) normalized Iterative Hard Thresholding (nIHT) [48] and 7) Stagewise weak Conjugate Gradient Pursuit (SWCGP) [152].[2]

[2]For the implementation of the other CS methods we used the MATLAB codes included in the packages: `http://sparselab.stanford.edu/`, `http://www.acm.caltech.edu/l1magic`, `http://ee.sharif.ir/~SLzero`, `http:`

The usual measure for the noise level in a corrupted signal is the Signal to Noise Ratio (SNR) defined as the ratio of the power of the observed data vector $\vec{g}$ to the power of the noise $\vec{\eta}$, which is expressed in dB's as follows:

$$\mathrm{SNR} = 10\log_{10}\Big(\frac{\|\vec{g}\|_2^2}{\|\vec{\eta}\|_2^2}\Big) \ . \tag{6.48}$$

However, a $S\alpha S$ distribution does not have finite second order statistics and therefore, SNR is not a valid distortion measure in the $S\alpha S$ case. As an alternative, we use a signal distortion measure based on fractional lower-order statistics, the so-called *Fractional-order SNR* (FSNR),

$$\begin{aligned} \mathrm{FSNR} &= 10\log_{10}\Big(\frac{\mathbb{E}\{|\vec{g}|^p\}}{\mathbb{E}\{|\vec{\eta}|^p\}}\Big) \\ &= 10\log_{10}\Big(\frac{(C(p_g,\alpha_g)\gamma_g)^{p_g}}{(C(p_\eta,\alpha_\eta)\gamma_\eta)^{p_\eta}}\Big) \ , \end{aligned} \tag{6.49}$$

where α_g, γ_g denote the characteristic exponent and the dispersion of $\vec{g}$, respectively, and p_g is the corresponding p-parameter depending on α_g (the same notation holds for $\vec{\eta}$). Notice that when the signal and noise components are jointly $S\alpha S$ ($\alpha_g = \alpha_\eta$), then

$$\mathrm{FSNR} = p \cdot 10\log_{10}\Big(\frac{\gamma_g}{\gamma_\eta}\Big) \ . \tag{6.50}$$

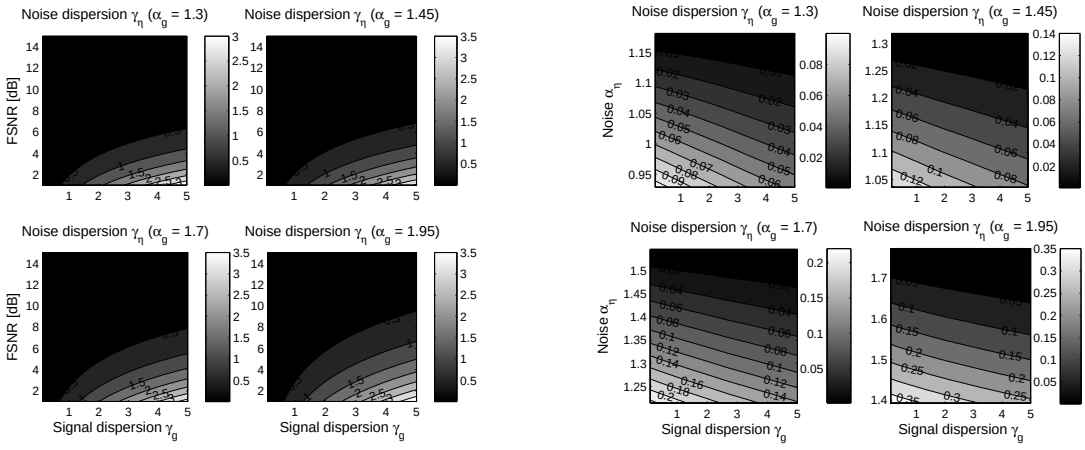
In the following simulations the noise dispersion is determined via Eq. (6.50) in the jointly $S\alpha S$ case, for a given pair (α_g, γ_g) and an FSNR value (in dB). Although the proposed $S\alpha S$-CS algorithm was developed under the assumption that the signal and noise components are jointly $S\alpha S$ however, the performance of the proposed method will be also evaluated in the case of additive $S\alpha S$ noise with a different characteristic exponent from that of the signal. In this case the noise dispersion is determined via Eq. (6.49) when $\alpha_g \neq \alpha_\eta$, for a given triplet $(\alpha_g, \alpha_\eta, \gamma_g)$ and an FSNR value (in dB).

Figures 6.3(a)-6.3(b) show contour lines for four different values of α_g illustrating the behavior of noise dispersion γ_η, as the signal dispersion γ_g varies in $[0.1, 5]$, in the jointly $S\alpha S$ case (Eq. (6.50)) and when the signal and noise components have different characteristic exponents (Eq. (6.49)), respectively. In the second case the FSNR value is fixed at 10 dB.

6.5.1 Performance evaluation of $S\alpha S$-CS (Algorithm 4)

In the first test case, we evaluate the performance of the proposed $S\alpha S$-CS algorithm for reconstructing simulated $S\alpha S$ signals corrupted by $S\alpha S$ noise with the same or different characteristic exponents. Notice once again that the proposed method has been developed under the assumption of a sparse vector with non-negative components. For this purpose, we generate vectors $\vec{x}$ of length $N = 512$ with zero components except for $L = 10$ randomly chosen positions whose values are drawn from a $S\alpha S$ distribution. Then, the non-negative sparse vector to be reconstructed is $\vec{w} = abs(\vec{x})$, where $abs(\vec{x}) = (|x_1|, \ldots, |x_N|)$. The value of α varies in the interval $[1.1, 2]$, while the dispersion γ_w is chosen from $[0.1, 5]$. Accordingly, following the discussion in Section 6.2.1 the entries of the measurement matrix $\mathbf{\Phi}$ are chosen from a $S\alpha S$ distribution with the same α as the sparse vector and dispersion 1. The noise dispersion γ_η is determined via Eq. (6.49) or Eq. (6.50) for a given pair (α_w, γ_w) or a triplet $(\alpha_w, \alpha_\eta, \gamma_w)$, respectively, and FSNR values (in dB) ranging in the interval $[5, 15]$. For a given α, the corresponding optimal value of p is given by interpolating the "Optimal q" columns of Table 3.1 as a function of α.

`//www.see.ed.ac.uk/~tblumens/sparsify/sparsify.html`.



(a) (b)

Figure 6.3: Noise dispersion contours: (a) Noise dispersion contours in the jointly $S\alpha S$ case, as a function of γ_g and FSNR, for $\alpha_g = 1.3, 1.45, 1.7, 1.95$, (b) Noise dispersion contours as a function of γ_g and α_η, for $\alpha_g = 1.3, 1.45, 1.7, 1.95$ and FSNR = 10 dB.

We also note that the subsequent performance evaluation is represented as an average over 100 Monte-Carlo runs.

Besides, the parameter ξ involved in the basis selection rule is set to 1.035 in order to accelerate the $S\alpha S$-CS approach by permitting the simultaneous selection of more than one basis vectors in each iteration. Figure 6.4 shows the relative reconstruction SNR value of the proposed $S\alpha S$-CS method in comparison with the values achieved by the other CS methods, where the relative reconstruction SNR is defined as follows,

$$\text{rSNR} = 10\log_{10}\left(\frac{\|\vec{w}\|_2^2}{\|\vec{w}-\widehat{\vec{w}}\|_2^2}\right), \tag{6.51}$$

where $\widehat{\vec{w}}$ is the reconstructed sparse vector. The rSNR is shown as a function of the number of CS measurements $M \in \{80:10:120\}$ for $\alpha = 1.3$, $\gamma_w = 0.7$ and FSNR = 10 dB (resulting in $\gamma_\eta = 0.018$). As we can see, the proposed CS algorithm outperforms all the other methods except for the LASSO technique, which follows closely $S\alpha S$-CS. However, notice that as M increases $S\alpha S$-CS outperforms LASSO, too. This can be justified by the fact that since the proposed method depends on the p-parameter, whose value is a function of α, then, by increasing the size M of the measurement vector $\vec{g}$ results in a more accurate estimation of α and thus, of p. In addition, we observe that for this specific heavy-tailed environment the L1EQ-PD and nIHT methods actually fail completely to reconstruct the sparse vector $\vec{w}$.

On the other hand, Figure 6.5 shows the CS ratio for each one of the selected CS methods, which we define as follows:

$$\text{CS ratio} = \frac{\text{number of CS measurements } M}{\text{number of non-zero components of } \vec{w}}, \tag{6.52}$$

where the number of non-zero components of $\vec{w}$ (sparsity) depends on the algorithm. The higher the CS ratio the higher the sparsity is for a fixed value of M. We can see that on average the proposed $S\alpha S$-CS method results in much higher CS ratios and thus, in much sparser solutions when compared with the other methods. Also notice that as M increases the proposed $S\alpha S$-CS algorithm converges to the true sparsity $L = 10 \simeq \frac{M}{\{\text{CS ratio}\}}$, in contrast to the other methods.

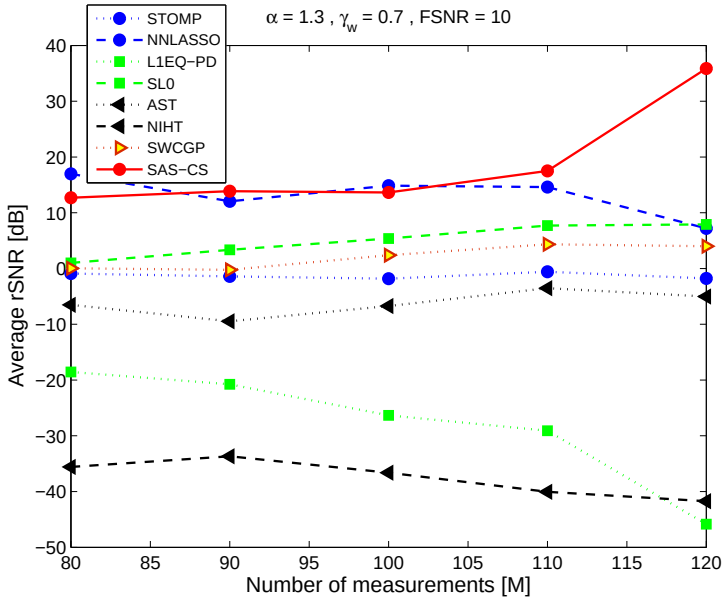


Figure 6.4: Average rSNR's as a function of the number of CS measurements M ($\alpha = 1.3$, $\gamma_w = 0.7$ and FSNR = 10 dB).

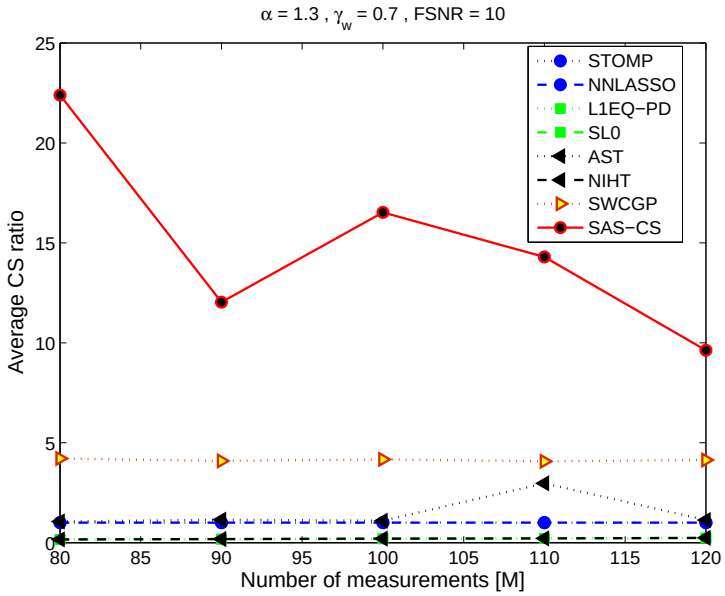


Figure 6.5: Average CS ratios of the CS-based recovery methods as a function of the number of CS measurements M ($\alpha = 1.3$, $\gamma_w = 0.7$ and FSNR = 10 dB).

Similarly, Figures 6.6-6.7 show the relative reconstruction SNR values and the corresponding CS ratios, respectively, of the proposed $S\alpha S$-CS method in comparison with the values achieved by the other CS methods, by setting $\alpha = 1.5$, $\gamma_w = 1$ and FSNR = 10 dB (resulting in $\gamma_\eta = 0.037$). As before, the proposed CS algorithm outperforms all the other methods as M increases, while also converging to the true sparsity $L = 10$.

Experimentation with other (α, γ_w, FSNR) triplets revealed a similar behavior as in Figures 6.4-6.7, in the sense that the LASSO method is the main competitor of the proposed $S\alpha S$-CS method. Thus, in the subsequent evaluation we primarily focus on comparing the proposed method with LASSO.

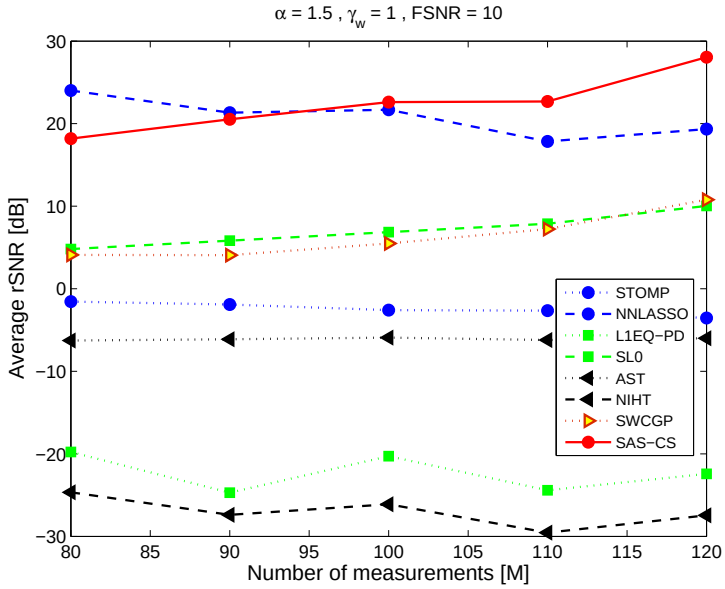


Figure 6.6: Average rSNR's as a function of the number of CS measurements M ($\alpha = 1.5$, $\gamma_w = 1$ and FSNR = 10 dB).

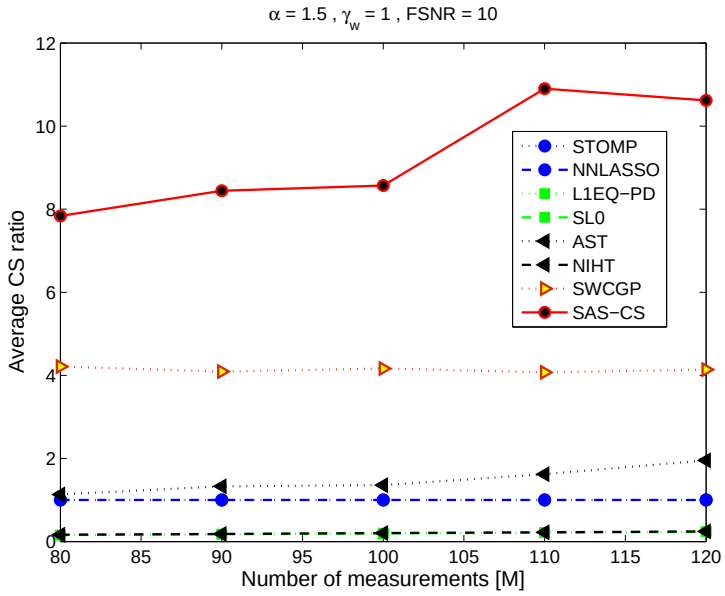


Figure 6.7: Average CS ratios of the CS-based recovery methods as a function of the number of CS measurements M ($\alpha = 1.5$, $\gamma_w = 1$ and FSNR = 10 dB).

Figure 6.8 on the next page shows the average reconstruction rSNR values for the $S\alpha S$-CS and the LASSO methods as a function of M, for $\gamma_w = 1$, FSNR = 10 dB and $\alpha = 1.4, 1.5, 1.7, 1.8$. We observe that for small α (that is, for highly impulsive signals) the proposed $S\alpha S$-CS method outperforms LASSO. However, as $\alpha \to 2$ (Gaussian assumption) then the LASSO results in a better reconstruction performance. On the other hand, Figure 6.9 shows that this increased performance comes at the cost of a significant increase in the number of basis functions used by LASSO (whose CS ratio is much smaller than the corresponding value of $S\alpha S$-CS). Notice also that again the $S\alpha S$-CS method tends to the true sparsity, $L = 10$, in contrast to LASSO.

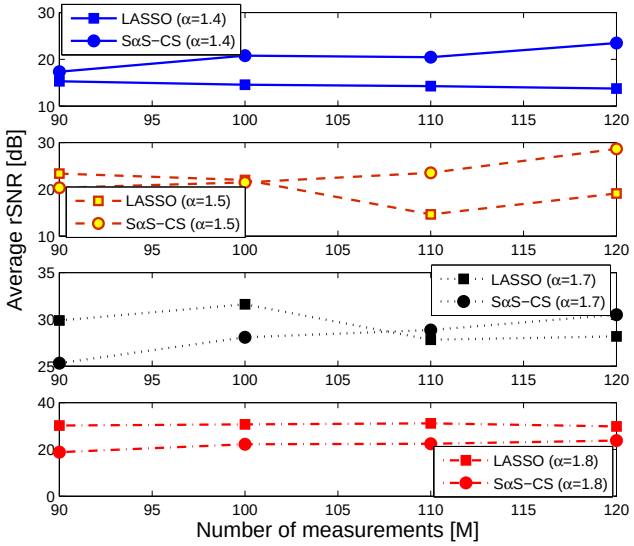


Figure 6.8: Average rSNR for $S\alpha S$-CS and LASSO as a function of the number of CS measurements M, for $\gamma_w = 1$, FSNR $= 10$ dB and $\alpha = 1.4, 1.5, 1.7, 1.8$.

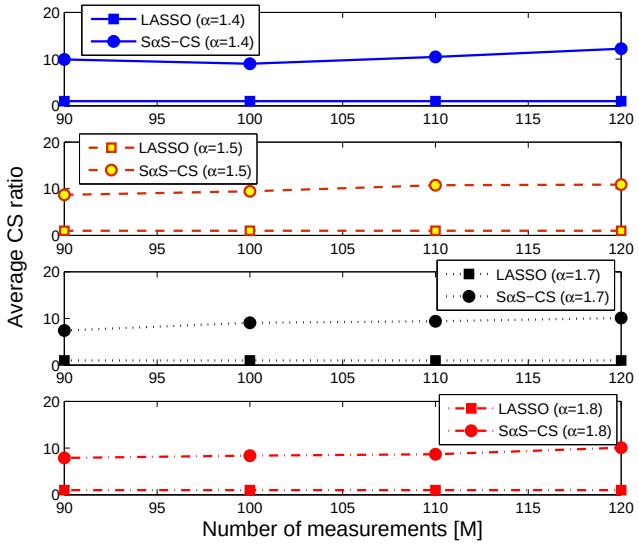
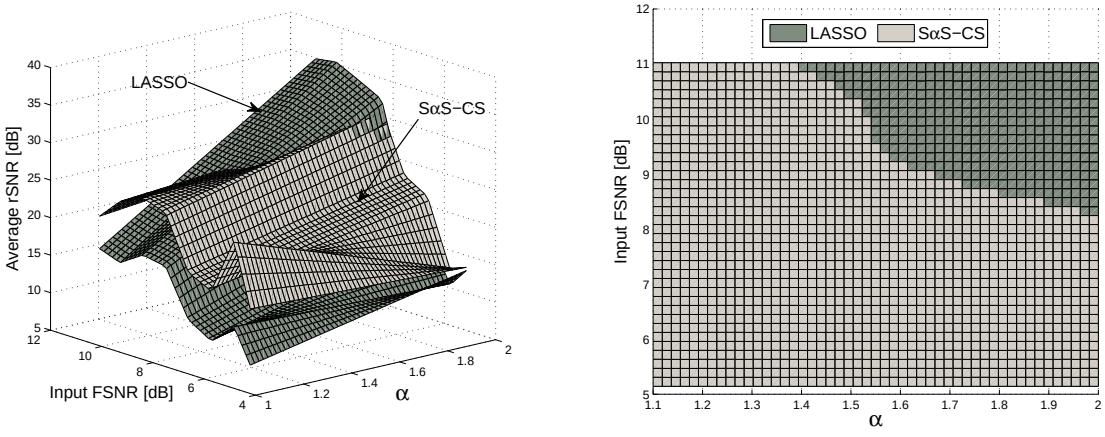


Figure 6.9: Average CS ratios for $S\alpha S$-CS and LASSO as a function of the number of CS measurements M, for $\gamma_w = 1$, FSNR $= 10$ dB and $\alpha = 1.4, 1.5, 1.7, 1.8$.

In order to examine the influence of the FSNR value on the reconstruction performance we plot in Figure 6.10(a) the reconstruction rSNR of the proposed $S\alpha S$-CS method in comparison with LASSO as a function of α and FSNR for $\gamma_w = 1$ and $M = 100$. For a better visualization, Figure 6.10(b) plots the same results from a top view. As it can be seen, $S\alpha S$-CS outperforms LASSO in the majority of (α, FSNR) pairs. Most importantly, $S\alpha S$-CS results in a better reconstruction performance as α is getting smaller (that is, as the signal model tends to be more impulsive and with heavier tails). Figure 6.11 shows the corresponding CS ratio values, illustrating again the ability of the $S\alpha S$-CS method to approach the true sparsity (especially as the FSNR value increases) in contrast to LASSO.



(a) (b)

Figure 6.10: (a) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and FSNR (in dB) for $\gamma_w = 1$ and $M = 100$, (b) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and FSNR (in dB) [Top view].

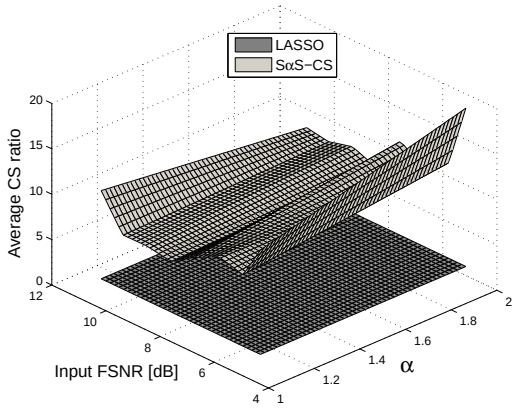
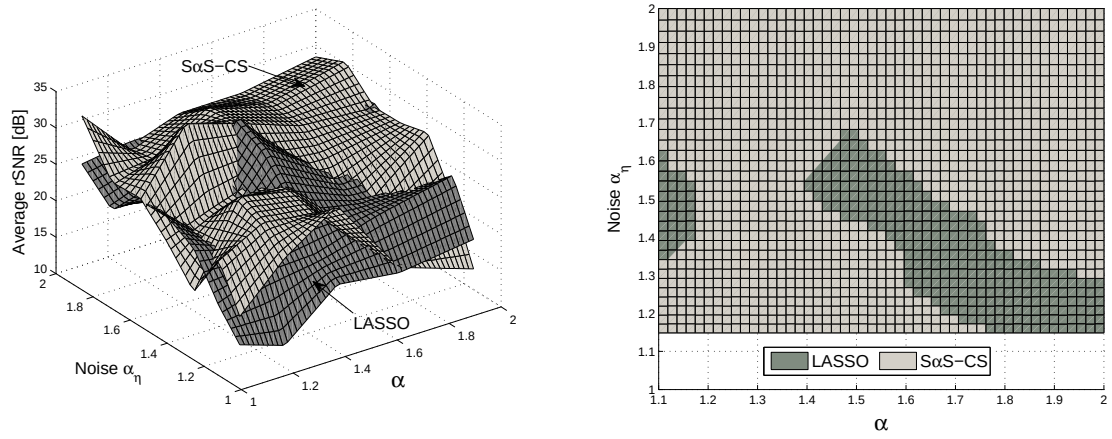


Figure 6.11: Average CS ratio for $S\alpha S$-CS and LASSO as a function of α and FSNR (in dB) for $\gamma_w = 1$ and $M = 100$.

Finally, as a last illustration, we evaluate the performance of the proposed $S\alpha S$-CS algorithm when the signal and noise components are not jointly $S\alpha S$, that is, when $\alpha_w \neq \alpha_\eta$. For this purpose, we plot in Figure 6.12(a) on the following page the reconstruction rSNR of the proposed $S\alpha S$-CS method in comparison with LASSO as a function of α (α_w) and α_η for $\gamma_w = 1$, $M = 100$ and FSNR = 8 dB. For a better visualization, Figure 6.12(b) plots the same results from a top view. As it can be seen, $S\alpha S$-CS outperforms LASSO in the majority of (α, α_η) pairs with the difference being more prominent for smaller values of α, justifying for one more time the validity of the proposed method in a truly impulsive environment. Figure 6.13 shows the corresponding CS ratio values, revealing again the ability of the $S\alpha S$-CS method to approach the true sparsity in contrast to LASSO.



(a) (b)

Figure 6.12: (a) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and α_η for $\gamma_w = 1$, $M = 100$ and FSNR = 8 dB, (b) Average rSNR for $S\alpha S$-CS and LASSO as a function of α and α_η [Top view].

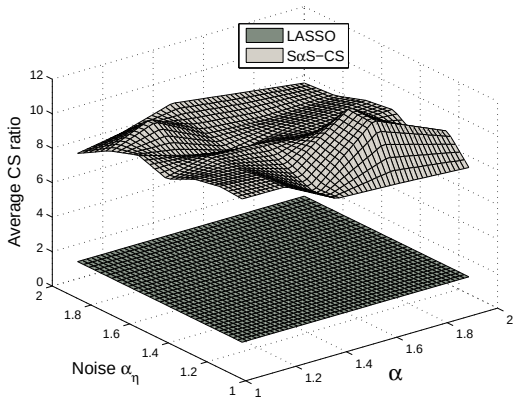


Figure 6.13: Average CS ratio for $S\alpha S$-CS and LASSO as a function of α and α_η for $\gamma_w = 1$, $M = 100$ and FSNR = 8 dB.

6.5.2 Performance evaluation of distributed $S\alpha S$-CS (Algorithm 5)

In the second test case we evaluate the performance of the proposed distributed implementation of $S\alpha S$-CS as described by Algorithm 5. For this purpose, a set of simulated $S\alpha S$ sparse vectors is generated with the specifications described in the first test case.

First, we validate the efficiency of the proposed FLOM-based Lagrangian function, given by Eq. (6.30), in capturing the significant basis functions (columns of $\mathbf{\Phi}$) to be activated for the estimation of the sparse vector $\vec{w}$, in contrast to the standard Lagrangian given by Eq. (6.25). We do so using simulated $S\alpha S$ signal ($\vec{w}$) and noise ($\vec{\eta}$) components with $\alpha \in [1.1, 2]$, $\gamma_w = 0.7$, FSNR $\in [5, 15]$ by repeating the process for each triplet $(\alpha, \gamma_w = 0.7, \text{FSNR})$ for 100 Monte-Carlo runs. Then, for each signal $\vec{w}$ the $S\alpha S$-CS reconstruction algorithm (Algorithm 4) is executed to estimate $\hat{\vec{w}}$, as well as the corresponding set of significant basis functions, whose indices are stored in a vector $\mathcal{T}_\mathcal{R}$.

Algorithm 5 is executed next using the standard and the FLOM-based Lagrangian function

for reconstructing the same sparse vector $\vec{w}$, resulting in the indices vectors $\mathcal{T}$ and $\mathcal{T}_{\mathcal{S}}$, respectively, containing the corresponding indices of the significant basis functions[3]. Algorithm 5 proceeds by setting $K = 15$, $M = 100$, $\beta = 1$, MaxIter $= 2N$ and $\varepsilon = 10^{-6}$. We also note that a different partition of $\mathbf{\Phi}$ (and $\vec{w}$) is created in each Monte-Carlo run, by assigning a different number of columns n_k to the k-th sensor ($k = 1, \ldots, K$). However, we take care of generating "balanced" partitions in the sense that all sensors obtain a similar number of columns of $\mathbf{\Phi}$.

Figure 6.14 on the next page shows the average percentage of successful retrievals of the significant basis functions, as expressed via the cardinalities of the intersections $\mathcal{T}_{\mathcal{R}} \cap \mathcal{T}$ and $\mathcal{T}_{\mathcal{R}} \cap \mathcal{T}_{\mathcal{S}}$ as a function of α and FSNR (in dB). It is clear that, on average, the standard Lagrangian function, which is based on second-order statistics, is able to retrieve less than half of the significant basis functions as estimated by the $S\alpha S$-CS method. On the other hand, the distributed $S\alpha S$-CS algorithm combined with the FLOM-based Lagrangian function has an 100% percentage of success in retrieving the significant basis functions given by its "centralized" ($S\alpha S$-CS) implementation.

Figures 6.15(a)-6.15(b) show the relative reconstruction SNR for the "centralized" $S\alpha S$-CS method (Algorithm 4) and its distributed extension (Algorithm 5), as a function of α and FSNR (in dB). First of all, we observe that the reduced dimensionality problem, resulting by implementing the distributed $S\alpha S$-CS method, which is then solved at the central node, yields exactly the same reconstruction performance with the solution of its "centralized" full dimensional counterpart. In addition, we can see that for both methods the reconstruction performance increases as the values of α and FSNR increase. The decrease of rSNR as $\alpha \to 1$ is related to the increased inaccuracy in estimating the characteristic exponent α using a measurement vector $\vec{g}$ of small size $M = 100$. This problem can be alleviated by increasing the number of measurements M. For instance, in practical applications employing time-series a significant amount of data is usually available and thus, one can increase M in order to enhance the estimation accuracy of α without violating the requirement of a CS-based method for a small set of measurements (M will be still small compared to the length of the original time-series).

6.6 Conclusions and future work

In this chapter, we described a distributed method for CS reconstruction based on a nonlinear programming framework with a potential application in a WSN. The sparse weight vector is modelled directly with a heavy-tailed distribution selected from the family of $S\alpha S$ distributions that enforces its sparsity. The experimental results revealed an increased reconstruction performance of highly impulsive vectors with non-negative components, while also achieving an increased sparsity, when compared with other state-of-the-art iterative greedy CS algorithms. Besides, we showed that the computational cost for acquiring and processing the sparse signal at each sensor is reduced significantly to satisfy the limitations of a WSN.

As future work, we are interested in exploiting the $S\alpha S$ model for developing a CS algorithm in a purely Bayesian framework. We expect that a probabilistic approach will provide further control on the sparsity of the weight vector and furthermore it will permit the optimal design of future CS measurements with the goal of reducing the uncertainty of the reconstructed signal, something which is not possible with the present norm-based iterative approach. In addition, we will extend the $S\alpha S$-CS method in the case of non-jointly $S\alpha S$ signal and noise components ($\alpha_w \neq \alpha_\eta$).

[3]We mention for clarification that when using the standard Lagrangian function the i-th column of $\mathbf{\Phi}$, $\vec{\phi}_i$, is considered to be significant if $|\vec{\phi}_i^T \vec{\lambda}| \geq 1$.

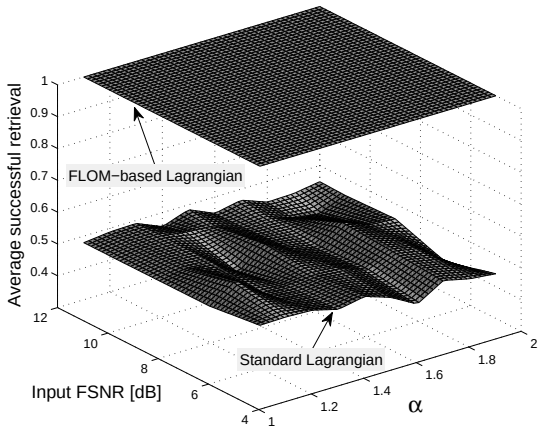
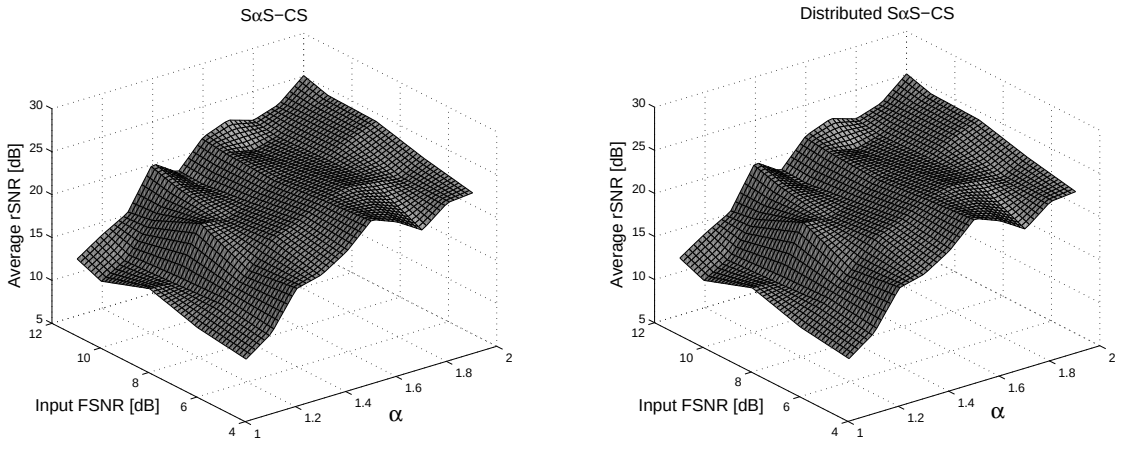


Figure 6.14: Average percentage of successful retrievals of the significant basis functions for the standard and FLOM-based Lagrangian function, as a function of α and FSNR (in dB).



(a) (b)

Figure 6.15: (a) Average reconstruction rSNR (in dB) for the "centralized" $S\alpha S$-CS method, as a function of α and FSNR (in dB), (b) Average reconstruction rSNR (in dB) for the distributed $S\alpha S$-CS method, as a function of α and FSNR (in dB).

Chapter

7

Conclusions and Future work

> The trouble with our times is that the future is not what it used to be.
>
> P. Valery (1871-1945)
> *French poet and essayist*

Concluding this thesis, we review in brief the main contributions and results presented in the previous chapters, while also providing directions for future work.

7.1 Thesis overview

The subject of this work was the development of CS algorithms for representing and reconstructing highly sparse signals in a Bayesian framework. Although the vast majority of the previous CS techniques was based on the solution of norm-based, constrained optimization problems, however, designing a purely probabilistic (Bayesian) method offers the critical advantage that it results not simply in a point estimate of the sparse vector, but it also provides confidence intervals, which be exploited for the design of adaptive future measurements with the goal of reducing the estimation uncertainty. In this framework, the prior belief for a sparse vector is expressed by modeling its prior distribution with an appropriate sparsity-enforcing distribution function.

The first, recently introduced, Bayesian CS methods were designed by modeling the prior distribution of the sparse vector using a hierarchical model in order to overcome the lack of closed-form expressions in the Bayesian inference. In *Chapter 4* we introduced a method for representing highly sparse signals by replacing the hierarchical model with a GSM, which is applied directly on the signal to be reconstructed enforcing its sparsity. The main advantage of the proposed method (BCS-GSM) is due to the scaling factor of the GSM model. This factor provides an additional degree of freedom by modeling the heavy-tailed behavior of the sparse signal and thus providing a more accurate representation of the true underlying statistics of the signal and/or the noise. The performance of BCS-GSM was evaluated on several sets of 1-D synthetic signals, as well as on real-world images (sparsified in the DWT domain). In both test cases, the proposed method either maintained the reconstruction performance of other recent state-of-the-art CS method, or it outperformed them. Most importantly, in all of the cases considered our method resulted in significantly increased sparse representations by employing much less significant basis vectors, expressed by the higher CS ratios.

There are several practical applications where multiple measurement vectors of the same original signal are available. For this purpose we extended BCS-GSM in order to take into account the multiple observations (MMV BCS-GSM). We showed that the proposed extension

is amenable to a distributed implementation using two different network topologies. We evaluated its performance by applying it on the problems of DOA estimation and image sparse representation. In both cases the results revealed an increased reconstruction performance of the distributed implementation, which is close to its SMV ("centralized") counterpart, namely the BCS-GSM method, as well as to other state-of-the-art CS methods, or even it outperforms them in several experimental scenarios. However, in all cases the distributed version resulted in significantly increased CS ratios, which is equivalent to representing a sparse signal using much fewer basis vectors. In addition, in the case of images MMV BCS-GSM revealed another remarkable property, namely the ability of acting as a diffusion operator resulting in smoother images with a higher degree of denoising.

The Gaussian assumption has been adopted in the vast majority of the previous CS methods, since the Bayesian inference results in simple, closed-form expressions. However, this assumption is far from being adequate in modeling the statistical characteristics of the signal and/or noise components in several real-world environments. In *Chapter 5*, we overcome this limitation of the Gaussian model by introducing a Bayesian CS method for recovering highly impulsive signals corrupted by additive impulsive noise. In particular, the highly impulsive nature is modelled by approximating the prior distribution of the signal and/or noise components with a Cauchy distribution, which is heavy-tailed and thus appropriate in representing such a behavior. Besides, the Bayesian inference is carried out by employing a suitable tree-structure, which results in an efficient implementation of the proposed method using closed-form expressions. The performance of the proposed CS algorithm was evaluated and compared with other CS methods based on a Gaussian assumption. The results revealed a significant improvement of the reconstruction quality when the signal and/or the noise is generated in a truly impulsive environment, while yielding much sparser solutions, in contrast to the other methods.

The Cauchy distribution employed in the previous method belongs to the family of $S\alpha S$ distributions, which are heavy-tailed and thus appropriate in modeling highly sparse signals. Despite this fact, the $S\alpha S$ distributions have not gained the interest of the research community due to the lack of closed-form density function (except for the Cauchy, and the Gaussian), as well as the lack of second-order moments. In *Chapter 6*, we introduced a novel CS algorithm, where the prior belief that the signal and/or noise are highly sparse (impulsive) is expressed by modeling their corresponding prior distributions with a member of the $S\alpha S$ family. First, we proposed a new measurement matrix, which is best adapted to the inherent statistical characteristics of the sparse signal, with its i.i.d. entries being drawn from a standard $S\alpha S$ distribution with the same characteristic exponent as the one estimated from the acquired sparse signal.

Then, we developed a novel iterative greedy algorithm for CS reconstruction of signals with non-negative components by solving a constrained optimization problem using duality theory and subgradients. For this purpose, a novel Lagrangian function was introduced which is best adapted to the case of $S\alpha S$ distributions by exploiting Fractional Lower-Order Moments instead of second-order statistics, which are not defined for this family of distributions. Moreover, we showed that the associated dual problem is separable and thus amenable to a distributed implementation. Motivated by this observation, we extended the $S\alpha S$-based CS method in a distributed fashion with a potential application in a WSN.

The experimental results revealed an increased reconstruction performance of highly impulsive vectors with non-negative components, while also achieving an increased sparsity when compared with other state-of-the-art iterative greedy CS algorithms. Besides, we showed that the computational cost for acquiring and processing the sparse signal at each sensor is reduced significantly to satisfy the limitations of a WSN.

7.2 Future work

In this section, we propose several research direction for improving and extending the CS algorithms introduced in the previous chapters with respect to the algorithmic part, as well as potential applications which could benefit from their increased performance.

First, the proposed GSM-based CS algorithm described in *Chapter 4* was developed without making any assumption for the prior distribution of the scaling factor. However, when this prior distribution is chosen from the family of positive α-Stable distributions, then the GSM model reduces to a sub-Gaussian model, which is expected to be more efficient in characterizing a highly sparse signal due to the additional degrees of freedom provided by its characteristic exponent and dispersion, resulting in a more accurate approximation of the true heavy-tailed statistics of the sparse signal. Besides, the distributed extension of the proposed BCS-GSM method was implemented using to distinct network topologies, namely a star-shaped and a ring-shaped one. As a future work, several clustering strategies could be examined in order to reduce the in-network communications, while maintaining a high performance in terms of the reconstruction quality and the achieved sparsity.

In *Chapter 5*, the proposed CS method was developed by exploiting the one-to-one correspondence of a sub-Gaussian vector with its underlying Gaussian part. As a direct extension, this method could be modified such as to employ directly the estimated covariation matrix (FLOM-based) instead of the associated covariance matrix (second-order statistics).

The $S\alpha S$-based CS algorithm presented in *Chapter 6* was developed under the assumption of jointly $S\alpha S$ signal and noise components. A direct generalization is the modification of this method in order to consider the case where the signal and the noise are drawn from $S\alpha S$ distributions with different characteristic exponents. The greedy algorithm described in this chapter does not provide the advantage for acquiring adaptively future measurements with the goal of reducing the estimation uncertainty, as the purely Bayesian methods do. As a future work, we intend in extending the proposed approach in a purely Bayesian framework. Besides, the distributed version was implemented in a simple star-shaped network topology. As a generalization, the proposed distributed method could be modified so as to adapt in several network topologies and clustering strategies with the goal of reducing further the communication and processing cost.

The performance of all the CS algorithms introduced in this thesis was evaluated primarily for the reconstruction of sparse signals (1-D or 2-D). However, there are several applications where an accurate reconstruction of the original signal is not necessary. Motivated by this observation, the highly sparse representations achieved by the proposed methods can be exploited in carrying out several signal processing tasks, such as detection, classification, and retrieval. This will be one of the main directions of our ongoing research.

Bibliography

[1] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. AIEE*, Vol. 47, pp. 617–644, Apr. 1928.

[2] C. E. Shannon, "Communication in the presence of noise," *Proc. Institute of Radio Engineers*, Vol. 37, No. 1, pp. 10–21, Jan. 1949.

[3] E. Candès, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, Vol. 52, No. 2, pp. 489–509, Feb. 2006.

[4] D. Donoho, "Compressed Sensing," *IEEE Trans. Inform. Theory*, Vol. 52, No. 4, pp. 1289–1306, Apr. 2006.

[5] E. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, Vol. 52, No. 12, pp. 5406–5425, Dec. 2006.

[6] S. Mallat, "A Wavelet Tour of Signal Processing," 2nd ed., New York: Academic Press, 1998.

[7] I. Daubechies, "Ten lectures on wavelets," SIAM, 1992.

[8] D. S. Taubman and M. W. Marcellin, "JPEG 2000: Image Compression Fundamentals, Standards and Practice," (Int. Series in Engineering and Computer Science), Norwell, MA: Kluwer, 2002.

[9] R. Baraniuk, M. Davenport, R. A. DeVore and M. B. Wakin, "A simple proof of the Restricted Isometry Property for random matrices," *Constructive Approximation*, Jan. 2008.

[10] M. Rudelson and R. Vershynin, "Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements," *40th Annual Conf. on Information Sciences and Systems (CISS 2006).*

[11] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Prob.*, Vol. 23, No. 3, pp. 969–985, 2007.

[12] R. Coifman, F. Geshwind and Y. Meyer, "Noiselets," *Appl. Comput. Harmon. Anal.*, Vol. 10, No. 1, pp. 27–44, 2001.

[13] E. Candès, J. Romberg and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, Vol. 59, No. 8, pp. 1207–1223, Aug. 2006.

[14] D. Baron, M. Wakin, M. Duarte, S. Sarvotham and R. Baraniuk, "Distributed compressed sensing," *submitted to IEEE Trans. Inform. Theory* [Online: `http://dsp.rice.edu/cs/DCS112005.pdf`].

[15] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM J. Sci. Statist. Comput.*, Vol. 7, No. 4, pp. 1307–1330, 1986.

[16] D. Donoho and P. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, Vol. 49, pp. 906–931, 1989.

[17] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, Vol. 47, No. 7, pp. 2845–2862, 2001.

[18] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Natl. Acad. Sci., USA* , Vol. 100, pp. 2197–2202, 2003.

[19] M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of $\mathbb{R}^N$ bases," *IEEE Trans. Inform. Theory*, Vol. 48, No. 9, pp. 2558–2567, 2002.

[20] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, Vol. 49, No. 12, pp. 3320–3325, 2003.

[21] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inform. Theory*, Vol. 50, No. 6, pp. 1341–1344, 2004.

[22] A. Gilbert, S. Muthukrishnan and M. Strauss, "Improved time bounds for near-optimal sparse Fourier representation," *Proc. of SPIE 5914 (Wavelets XI)*, ed. by M. Papadakis, A. Laine and M. Unser, 2005.

[23] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B*, Vol. 58, No. 1, pp. 267–288, 1996.

[24] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, Vol. 35, No. 6, pp. 2313–2351, 2007.

[25] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, Vol. 52, No. 9, pp. 4036–4048, Sept. 2006.

[26] Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," *Signal Proc.*, Vol. 86, No. 3, pp. 549–571, Mar. 2006.

[27] J. Tropp, "Just relax: Convex programming methods for identifying sparse signal," *IEEE Trans. Inform. Theory*, Vol. 51, No. 3, pp. 1030–1051, 2006.

[28] I. Gorodnitsky and B. Rao, "Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm," *IEEE Trans. Signal Proc.*, Vol. 45, pp. 600–616, Mar. 1997.

[29] S. J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Sel. Topics in Sig. Proc.*, Vol. 1, No. 4, pp. 606–617, Dec. 2007.

[30] E. Candès and J. Romberg, "ℓ_1-Magic: A Collection of MATLAB Routines for Solving the Convex Optimization Programs Central to Compressive Sampling," 2006 [Online: `www.acm.caltech.edu/l1magic/`].

[31] E. Hale, W. Yin and Y. Zhang, "A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing," (Preprint, 2007) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[32] D. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," . *IEEE Trans. Inform. Theory*, Vol. 54, No. 11, pp. 4789–4812, 2008.

[33] D. Malioutov, M. Hetin and A. Willsky, "Homotopy continuation for sparse signal representation," *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Proc. (ICASSP'05)*, Philadelphia, PA, Vol. 5, pp. 733–736, 2005.

[34] M. Figueiredo, R. Nowak and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics in Sig. Proc.: Special Issue on Convex Optimization Methods for Signal Processing*, Vol. 1, No, 4, pp. 586–598, 2007.

[35] I. Daubechies, M. Fornasier and I. Loris, "Accelerated projected gradient method for linear inverse problems with sparsity constraints," (Preprint, 2007) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[36] T. Blumensath and M. Davies, "Gradient pursuits," *IEEE Trans. Signal Proc.*, Vol. 56, No. 6, pp. 2370–2382, June 2008.

[37] S. Chen, D. Donoho and M. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM Rev.*, Vol. 43, No. 1, pp. 129–159, 2001.

[38] J. Tropp and A. Gilbert, "Signal recovery from random measurements via Orthogonal Matching Pursuit," *IEEE Trans. Inform. Theory*, Vol. 53, No. 12, pp. 4655–4666, Dec. 2007.

[39] D. Donoho, Y. Tsaig, I. Drori and J.-L. Starck, "Sparse solution of underdetermined linear equations by Stagewise Orthogonal Matching Pursuit (StOMP)," (Preprint, 2007) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[40] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via Regularized Orthogonal Matching Pursuit," *Foundations of Comput. Mathematics*, Vol. 9, No. 3, pp. 317–334, 2007.

[41] V. Saligrama and M. Zhao, "Thresholded basis pursuit: Quantizing linear programming solutions for optimal support recovery and approximation in compressed sensing," (Preprint, 2008) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[42] Y. Eldar and H. Blcskei, "Block-sparsity: Coherence and efficient recovery," *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Proc. (ICASSP'09)*, Taipei, Taiwan, Apr. 19–24, 2009.

[43] Y. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," *submitted to IEEE Trans. Inform. Theory*, 2008.

[44] M. Stojnic, F. Parvaresh and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. on Signal Proc.*, Vol. 57, No. 8, pp. 3075–3085, 2009.

[45] M. Fornasier and H. Rauhut, "Iterative thresholding algorithms," *Appl. Comput. Harmon. Anal.*, Vol. 25, No. 2, pp. 187–208, 2008.

[46] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximations," *J. of Fourier Anal. and Appl.*, Vol. 14, No. 5, pp. 629–654, Dec. 2008.

[47] J. Bioucas-Dias and M. Figueiredo, "A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Proc.*, Vol. 16, No. 12, pp. 2992–3004, Dec. 2007.

[48] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *to appear in Applied & Comp. Harm. Anal.* [Online: `http://www.dsp.ece.rice.edu/cs/`].

[49] M. Smith and R. Kohn, "Nonparametric regression using Bayesian variable selection," *J. Econometrics*, Vol. 75, pp. 317–343, 1996.

[50] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learning Res.*, Vol. 1, pp. 211–244, 2001.

[51] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selectionn," *IEEE Trans. Signal Proc.*, Vol. 52, pp. 2153–2164, 2004.

[52] D. Wipf, J. Palmer and B. Rao, "Perspectives on sparse Bayesian learning," *Advances in Neural Inf. Proc. Systems (NIPS)*, Vancouver, Canada, Vol. 16, Dec. 2004.

[53] N. Vaswani, "Kalman filtered compressed sensing," *IEEE Int. Conf. on Image Proc. (ICIP)*, San Diego, California, Oct. 2008.

[54] A. Carmi, P. Gurfil and D. Kanevsky, "A simple method for sparse signal recovery from noisy observations using Kalman filtering," *IBM Res. Report* [RC24709], Dec. 2008.

[55] S. Ji, Y. Xue and L. Carin, "Bayesian Compressive Sensing," *IEEE Trans. on Signal Proc.*, Vol. 56, No. 6, pp. 2346–2356, June 2008.

[56] S. Ji and L. Carin, "Bayesian compressive sensing and projection optimization," *Proc. 24th Int. Conf. Machine Learning (ICML)*, pp. 377–384, 2007.

[57] M. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," *Proc. 25th Intern. Conf. on Machine Learning (ICML)*, pp. 912–919, 2008.

[58] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," in *Advances in Neural Inf. Proc. Systems (NIPS 14)*, 2002.

[59] M. Tipping and A. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," *in Proc. 9th Int. Workshop on Artificial Intell. and Statistics*, C. Bishop and B. Frey Eds., 2003.

[60] P. Schniter, Lee C. Potter and J. Ziniel, "Fast Bayesian Matching Pursuit," *Proc. Workshop on Inform. Theory & Applications*, La Jolla, CA, Jan. 2008.

[61] D. Baron, S. Sarvoham and R. Baraniuk, "Bayesian compressive sensing via belief propagation," (Preprint, 2008) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[62] M. Mishali and Y. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. on Signal Proc.*, Vol. 56, No. 10, pp. 4692–4702, 2008.

[63] S. Cotter, B. Rao, K. Engan and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Proc.*, Vol. 53, No. 7, pp. 2477–2488, July 2005.

[64] J. Tropp, A. Gilbert and M. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *EURASIP J. Signal Processing*, Vol. 86, pp. 572–588, Apr. 2006.

[65] J. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *EURASIP J. Signal Processing*, Vol. 86, pp. 589–602, Apr. 2006.

[66] S. Ji, D. Dunson and L. Carin, "Multi-task compressive sensing," (Preprint, 2007) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[67] Y. Qi, D. Liu, D. Dunson and L. Carin, "Bayesian multi-task compressive sensing with dirichlet process priors," (Preprint, 2008) [Online: `http://www.dsp.ece.rice.edu/cs/`].

[68] D. Estrin, D. Culler, K. Pister and G. Sukhatme, "Connecting the physical world with pervasive networks," *IEEE Pervasive Computing*, Vol. 1, No. 1, pp. 59–69, 2002.

[69] T. Cover and J. Thomas, "Elements of Information Theory," John Wiley and Sons, New York, 1991.

[70] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, Vol. 19, pp. 471–480, July 1973.

[71] Z. Xiong, A. Liveris and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Mag.*, Vol. 21, pp. 80–94, Sept. 2004.

[72] W. Bajwa, J. Haupt, A. Sayeed and R. Nowak, "Compressive wireless sensing," *Proc. Int. Conf. on Information Proc. in Sensor Networks (IPSN)*, Nashville, Tennessee, April 2006.

[73] M. Duarte, S. Sarvotham, D. Baron, M. Wakin and R. Baraniuk, "Distributed compressed sensing of jointly sparse signals," *Proc. of the Asilomar Conf. on Signals, Systems and Computers*, Oct. 30–Nov. 2, 2005, Pacific Grove, CA.

[74] M. Duarte, M. Wakin, D. Baron and R. Baraniuk, "Universal distributed sensing via random projections," *Proc. Int. Conf. on Information Proc. in Sensor Networks (IPSN)*, Nashville, Tennessee, April 2006.

[75] W. Wang, M. Garofalakis and K. Ramchandran, "Distributed sparse random projections for refinable approximation," *Proc. Int. Conf. on Information Proc. in Sensor Networks (IPSN)*, Cambridge, Massachusetts, April 2007.

[76] R. Wagner, R. Baraniuk, S. Du, D. Johnson and A. Cohen, "An architecture for distributed wavelet analysis and processing in sensor networks," *Proc. Int. Conf. on Information Proc. in Sensor Networks (IPSN)*, Nashville, Tennessee, April 2006.

[77] M. Crovella and E. Kolaczyk, "Graph wavelets for spatial traffic analysis," *Proc. IEEE Infocom*, Vol. 3, pp. 1848–1857, San Francisco, CA, 2003.

[78] R. Coifman and M. Maggioni, "Diffusion wavelets," *Appl. Comput. Harmon. Anal.*, Vol. 21, No. 1, pp. 53–94, July 2006.

[79] M. Duarte, M. Davenport, M. Wakin and R. Baraniuk, "Sparse signal detection from incoherent projections," *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Toulouse, France, May 2006.

[80] J. Haupt and R. Nowak, "Compressive sampling for signal detection," *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Honolulu, Hawaii, April 2007.

[81] J. Haupt, R. Castro, R. Nowak, G. Fudge and A. Yeh, "Compressive sampling for signal classification," *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, Oct. 2006.

[82] M. Duarte, M. Davenport, M. Wakin, J. Laska, D. Takhar, K. Kelly and R. Baraniuk, "Multiscale random projections for compressive classification," *IEEE Conf. on Image Proc. (ICIP)*, San Antonio, Texas, Sept. 2007.

[83] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly and R. Baraniuk, "The smashed filter for compressive classification and target recognition," *Computational Imaging V at SPIE Electronic Imaging*, San Jose, California, Jan. 2007.

[84] V. Cevher, M. Duarte and R. Baraniuk, "Distributed target localization via spatial sparsity," *European Signal Proc. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[85] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Mag.*, Vol. 25, No. 2, pp. 83–91, Mar. 2008.

[86] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly and R. Baraniuk, "Compressive imaging for video representation and coding," *Proc. Picture Coding Symposium (PCS)*, Beijing, China, April 2006.

[87] V. Stankovic, L. Stankovic and S. Cheng, "Compressive video sampling," *European Signal Proc. Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[88] M. Lustig, D. Donoho and J. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, Vol. 58, No. 6, pp. 1182–1195, Dec. 2007.

[89] H. Jung, K. Sung, K. Nayak, E. Y. Kim and J. C. Ye, "k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI," *to appear in Magnetic Resonance in Medicine*, 2008.

[90] J. Bobin, J.-L. Starck and R. Ottensamer, "Compressed Sensing in Astronomy," *IEEE J. Sel. Topics in Sig. Proc.*, Vol. 2, No. 5, pp. 718–726, Oct. 2008.

[91] P. Stoica and A. Nehorai, "MUSIC, Max. Likelihood, and Cramer-Rao Bound," *IEEE Trans. Acoustics, Speech & Sig. Proc.*, Vol. 37, No. 5, pp. 720–741, 1989.

[92] H. Cox, R. Zeskind and M. Owen, "Robust Adaptive Beamforming," *IEEE Trans. Acoustics, Speech & Sig. Proc.* Vol. 35, No. 10, pp. 1365–1377, 1987.

[93] S. Cotter and B. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. on Communications*, Vol. 50, No. 3, Mar. 2002.

[94] W. Bajwa, A. Sayeed and R. Nowak, "Compressed sensing of wireless channels in time, frequency, and space," *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, Oct. 2008.

[95] J. Laska, S. Kirolos, Y. Massoud, R. Baraniuk, A. Gilbert, M. Iwen and M. Strauss, "Random sampling for analog-to-information conversion of wideband signals," *IEEE Dallas Circuits and Systems Workshop (DCAS)*, Dallas, Texas, 2006.

[96] T. Ragheb, S. Kirolos, J. Laska, A. Gilbert, M. Strauss, R. Baraniuk and Y. Massoud, "Implementation models for analog-to-information conversion via random sampling," *Midwest Symposium on Circuits and Systems (MWSCAS)*, 2007.

[97] C. K. Chui, "An Introduction to Wavelets," Academic Press, 1992.

[98] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," 2nd Edition, Prentice Hall, 2002.

[99] J. Huang, "Study on the correlation properties of wavelet transform coefficients and the applications in a neural network-based hybrid image coding system," *Proc. CISST'03*, Las Vegas, USA, June 22-26, 2003.

[100] P. Lévy, "Calcul des Probabilités," Paris: Gauthier-Villars, 1925.

[101] S. Cambanis, G. Samorodnitsky and M. Taqqu, "Stable Processes and Related Topics," Birkhauser, Boston, 1991.

[102] G. Samorodnitsky and M. Taqqu, "Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance," Chapman & Hall, New York, 1994.

[103] C. Nikias and M. Shao, "Signal Processing with Alpha-Stable Distributions and Applications," John Wiley & Sons, New York, 1995.

[104] S. Cambanis and G. Miller, "Linear problems in pth order and stable processes," *SIAM J. Appl. Math.*, Vol. 41, pp. 43–69, 1981.

[105] S. Cambanis, C. Hardin and A. Weron, "Ergodic Properties of Stationary Stable Processes," *Stochastic Processes and their Appl.*, Vol. 24, pp. 1–18, 1987.

[106] W. DuMouchel, "Stable Distributions in Statistical Inference," Dept. of Statistics, Yale University, 1971.

[107] M. Shao and C. Nikias, "On Symmetric Stable Models for Impulsive Noise," USC-SIPI-231, Univ. of Southern California, Feb. 1993.

[108] G. Tsihrintzis and C. Nikias, "Robust Signal Detection and Pattern Classification using Fractional Lower-Order Statistics: Application to Radar," *IEEE Signal Proc. Society, Signal Proc. Workshop*, Washington, DC, Mar. 24–25, 1995.

[109] W. Willinger, M. Taqqu, W. Leland and D. Wilson, "Self-similarity in high-speed packet traffic: Analysis and modeling of ethernet traffic measurements," *Stat. Science*, pp. 67–85, Vol. 10, 1995.

[110] P. Georgiou, P. Tsakalides and C. Kyriakakis, "Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise," *IEEE Trans. Multimedia*, Vol. 1, pp. 291–301, Sep. 1999.

[111] P. Tsakalides, P. Reveliotis and C. Nikias, "Scalar quantization of heavy-tailed signals," *IEE Proc.-Vision, Imag. Sign. Proc.*, Vol. 147, No. 5, pp. 475–484, Oct. 2000.

[112] J. Nolan, "Parameterizations and modes of stable distributions," *Statistics & Probability Letters*, No. 38, pp. 187–195, 1998.

[113] J. Nolan, "Numerical calculation of stable densities and distribution functions," *Commun. Statist.-Stochastic Models*, Vol. 13, pp. 759–774, 1997.

[114] E. Fama and R. Roll, "Parameter estimates for symmetric stable distributions," *J. Amer. Statist. Assoc.*, Vol. 66, pp. 331–338, June 1971.

[115] A. Paulson, E. Holcomb and R. Leitch, "The estimation of the parameters of the stable laws," *Biometrika*, Vol. 62, pp. 163-170, 1975.

[116] I. Koutrouvelis, "An iterative procedure for the estimation of the parameters of stable laws," *Commun. Statist.-Simul.*, Vol. 10, No. 1, pp. 17–28, 1981.

[117] P. Georgiou and C. Kyriakakis, "Maximum likelihood parameter estimation under impulsive conditions, a sub-Gaussian signal approach," *IEEE Trans. Signal Proc.*, Vol. 86, No. 10, pp. 3061–3075, 2006.

[118] M. Schilder, "Some structure theorems for the symmetric stable laws," *The Annals of Mathematical Statistics*, Vol. 41, pp. 412–421, 1970.

[119] M. Shao and C. Nikias, "Signal processing with fractional lower order moments: Stable processes and their applications," *Proc. IEEE*, Vol. 81, pp. 986–1010, July 1993.

[120] G. Tzagkarakis, "Content-based Image Retrieval via Alpha-Stable Modeling of Texture Information," *M.Sc. Thesis*, Dept. of Computer Science, Univ. of Crete, Nov. 2004.

[121] J. Chambers, W. Cleveland, B. Kleiner and P. Tukey, "Graphical Methods for Data Analysis," Wadsworth, 1983.

[122] M. Wakin, "A Manifold Lifting Algorithm for Multi-View Compressive Imaging," *Proc. Picture Coding Symposium (PCS 2009)*, Chicago, Illinois, May 2009.

[123] J. Chen and X. Huo, "Sparse representations for multiple measurement vectors (MMV) in an overcomplete dictionary," *in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, pp. 257–260, Mar. 2005.

[124] D. Wipf and B. Rao, "An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem," *IEEE Trans. Signal Proc.*, Vol. 55, No. 7, pp. 3704–3716, July 2007.

[125] J. Davis and I. Dhillon, "Differential Entropic Clustering of Multivariate Gaussians," *in Neural Inform. Proc. Systems (NIPS'06)*, 2006.

[126] M. Taroudakis, G. Tzagkarakis and P. Tsakalides, "Classification of shallow-water acoustic signals via alpha-Stable modeling of the one-dimensional wavelet coefficients," *J. Acoust. Soc. Am. (JASA)*, Vol. 119, No. 3, pp. 1396–1405, Mar. 2006.

[127] P. Tsakalides and C. Nikias, "Space-Time Adaptive Processing in Stable Impulsive Interference," *in Proc. 31st Annual Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 3–6, 1997, pp. 389–393.

[128] P. Georgiou, P. Tsakalides and C. Kyriakakis, "Alpha-Stable Robust Modeling of Background Noise for Enhanced Sound Source Localization," *in Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, Phoenix, AZ, Mar. 15-19, 1999, pp. 3085–3088.

[129] A. Achim, P. Tsakalides and A. Bezerianos, "SAR image denoising via Bayesian wavelet shrinkage based on heavy-tailed modeling," *IEEE Trans. Geosc. and Rem. Sens.*, Vol. 41, No. 8, pp. 1773–1784, Aug. 2003.

[130] P. Tsakalides and C. Nikias, "Maximum likelihood localization of sources in noise modeled as a Cauchy process," *Proc. IEEE Milit. Comm. Conf. (MILCOM 1994)*, Fort Monmouth, NJ, Vol. 2, pp. 613–617, Oct. 1994.

[131] R. Raspanti, P. Tsakalides, C. Nikias and E. Del Re, "Cramer-Rao bounds for target angle and doppler estimation for airborne radar in Cauchy interference," *8th IEEE Sig. Proc. Workshop on Stat. Sig. and Array Proc.*, Corfu, Greece, June 24–26, 1996.

[132] M. Clyde and E. George, "Model uncertainty," *Statist. Sci.*, Vol. 19, No. 1, pp. 81–94, 2004.

[133] A. Lapidoth and S. Moser, "Capacity Bounds via Duality with Applications to Multiple-Antenna Systems on Flat Fading Channels," *IEEE Trans. on Information Theory*, Vol. 49, No. 10, pp. 2426–2467, Oct. 2003.

[134] S. Grgic, K. Kers and M. Grgic, "Image compression using wavelets", *Proc. of the IEEE Int. Symp. on Industrial Elec.*, Vol. 1, pp.99–104, 1999.

[135] M. Rabbat, J. Haupt, A. Singh and R. Nowak, "Decentralized compression and predistribution via random gossiping," *in Proc. of IPSN'06*, April 19–21, 2006, Nashville, TN, USA.

[136] E. Candès, M. Wakin and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Tech. rep.*, California Institute of Technology, 2007.

[137] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Sig. Proc.*, Vol. 41, No. 12, pp. 3397–3415, 1993.

[138] Y. Pati, R. Rezaifar and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proc.* 27^{th} *Asilomar Conf. on Sig., Systems & Comp.*, Nov. 1993.

[139] M. Bouvet and S. Schwartz, "Comparison of adaptive and robust receivers for signal detection in ambient underwater noise," *IEEE Trans. on Acoustic, Speech, Sig. Proc.*, Vol. 37, pp. 621–626, May 1989.

[140] P. Tsakalides, F. Trinci and C. Nikias, "Radar CFAR Thresholding in Heavy-Tailed Clutter Environments, *Proc.* 32^{nd} *Asilomar Conf. on Sig., Systems & Comp.*, Pacific Grove, CA, Nov. 1–4, 1998.

[141] D. Middleton, "Statistical, physical models of electromagnetic interference," *IEEE Trans. on Electr. Compat.*, Vol. 19, No. 3 pp. 106–127, 1977.

[142] P. Glasserman, "Portfolio Value-at-Risk with Heavy-Tailed Risk Factors," *Mathematical Finance*, Vol. 12, No. 3, pp. 239–269, July 2002.

[143] S. Babacan, R. Molina and A. Katsaggelos, "Fast Bayesian Compressive Sensing using Laplace Priors," *Proc. IEEE Int. Conf. on Acoust., Speech & Sig. Proc. (ICASSP'09)*, Taipei, Taiwan, April 2009.

[144] G. Tzagkarakis and P. Tsakalides, "Bayesian Compressed Sensing of a Highly Impulsive Signal in Heavy-Tailed Noise using a Multivariate Cauchy Prior," *Proc.* 17^{th} *European Sig. Proc. Conf. (EUSIPCO'09)*, Glasgow, Scotland, Aug. 24–28, 2009.

[145] G. Tzagkarakis and P. Tsakalides, "Bayesian Compressed Sensing Imaging using a Gaussian Scale Mixture," *submitted in ICASSP'2010* [Online: `http://www.csd.uoc.gr/~gtzag/publications/conference/2010-ICASSP_GSM.pdf`].

[146] D. Malioutov, M. Cetin and A. Willsky, "Optimal sparse representations in general overcomplete bases," *Proc. IEEE Int. Conf. on Acoust., Speech & Sig. Proc. (ICASSP'04)*, Montreal, Canada, May 2004.

[147] S. Boyd and L. Vandenberghe, "Convex Optimization", Cambridge University Press, 2004.

[148] D. Bertsekas, "Nonlinear Programming", Athena Scientific, 2^{nd} Edition, 1999.

[149] L. He, H. Chen and L. Carin, "Tree-Structured Compressive Sensing with Variational Bayesian Analysis," *subm. to IEEE Signal Proc. Letters*, 2009.

[150] H. Mohimani, M. Babaie-Zadeh and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed ℓ-0 norm," *IEEE Trans. Signal Proc.*, Vol. 57, No. 1, pp. 289-301, Jan. 2009.

[151] S. Cabrera, R. Dominguez, J. Rosiles, J. Vega-Pineda, " Variable-p affine scaling transformation algorithms for improved compressive sensing," *Proc. Sensor, Sig. & Info. Proc. Workshop (SenSIP)*, Sedona, Arizona, May 2008.

[152] T. Blumensath and M. Davies, "Stagewise Weak Gradient Pursuits," *to appear in IEEE Trans. Signal Proc.* [Preprint].