l0 Sparse signal processing and model selection with applications

Author:
Seneviratne, Seneviratne

Publication Date:
2012

DOI:
https://doi.org/10.26190/unsworks/15967

l_0 Sparse Signal Processing and Model Selection with Applications

Akila J. Seneviratne

School of Electrical Engineering and Telecommunications
University of New South Wales

Supervised by Professor Victor Solo

A thesis submitted for the degree of
Doctor of Philosophy (PhD)
2009-2012

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Seneviratne

First name: Seneviratne Other name/s: Mudiyanselage Akila Jayani

Abbreviation for degree as given in the University calendar: PhD

School: School of Electrical Engineering and Telecommunications Faculty: Faculty of Engineering

Title: l_0 Sparse Signal Processing and Model Selection with Applications

Abstract 350 words maximum: (PLEASE TYPE)

Sparse signal processing has far-reaching applications including compressed sensing, media compression/denoising/deblurring, microarray analysis and medical imaging. The main reason for its popularity is that many signals have a sparse representation given that the basis is suitably selected. However the difficulty lies in developing an efficient method of recovering such a representation.

To this aim, two efficient sparse signal recovery algorithms are developed in the first part of this thesis. The first method is based on direct minimization of the l_0 norm via cyclic descent, which is called the L0LS-CD (l_0 penalized least squares via cyclic descent) algorithm. The other method minimizes smooth approximations of sparsity measures including those of the l_0 norm via the majorization minimization (MM) technique, which is called the QC (quadratic concave) algorithm.

The L0LS-CD algorithm is developed further by extending it to its multivariate (V-L0LS-CD (vector L0LS-CD)) and group (gL0LS-CD (group L0LS-CD)) regression variants. Computational speed-ups to the basic cyclic descent algorithm are discussed and a greedy version of L0LS-CD is developed. Stability of these algorithms is analyzed and the impact of the penalty parameter and proper initialization on the algorithm performance are highlighted. A suitable method for performance comparison of sparse approximating algorithms in the presence of noise is established. Simulations compare L0LS-CD and V-L0LS-CD with a range of alternatives on under-determined as well as over-determined systems.

The QC algorithm is applicable to a class of penalties that are neither convex nor concave but have what we call the quadratic concave property. Convergence proofs of this algorithm are presented and it is compared with the Newton algorithm, concave convex (CC) procedure, as well as with the class of proximity algorithms. Simulations focus on the smooth approximations of the l_0 norm and compare them with other l_0 denoising algorithms.

Next, two applications of sparse modeling are considered. In the first application the L0LS-CD algorithm is extended to recover a sparse transfer function in the presence of coloured noise. The second uses gL0LS-CD to recover the topology of a sparsely connected network of dynamic systems. Both applications use Laguerre basis functions for model expansion.

The role of model selection in sparse signal processing is widely neglected in literature. The tuning/penalty parameter of a sparse approximating problem should be selected using a model selection criterion which minimizes a desired discrepancy measure. Compared to the commonly used model selection methods, the SURE (Stein's unbiased risk estimator) estimator stands out as one which does not suffer from the limitations of other methods. Most model selection criterion are developed based on signal or prediction mean squared error. The last section of this thesis develops an SURE criterion instead for parameter mean square error and applies this result to l_1 penalized least squares problem with grouped variables. Simulations based on topology identification of a sparse network are presented to illustrate and compare with alternative model selection criteria.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

Akila ..... Signature

Witness ..... Witness

14.02.2013 ..... Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged'

Signed .....

Date 14.02.2013

I lovingly dedicate this thesis to my husband Indika, my parents, sister and brother, for their endless love, support and encouragement.

Acknowledgements

It is impossible to overstate my gratitude to my supervisor Professor Victor Solo. I have been extremely fortunate to receive guidance from such an experienced, knowledgable and inspirational individual who has supported me throughout my thesis with patience and understanding.

I am eternally indebted to my loving husband, Indika, who helped me to be positive through many difficult times. I appreciate the sacrifices he made for me and my studies during our lives in two different countries.

I am sincerely grateful to my parents, sister and brother who always believed in me even when I did not. Their kind loving words never failed to lift my spirits.

Next I would like to thank my dear friends Ms. Mitra Bahadorian and Dr. Borislav Savkovic, who has been my friends since the commencement of my Ph.D. studies. I am so privileged to have such generous, selfless, caring friends and I do not know how I would have made through these years without them.

I would also like to extend my gratitude to all the other friends that I made since I came to study at UNSW. I am fortunate to have such caring friends and I will always reflect back fondly at the times we had together.

This thesis was partly funded by the International Telecommunication Union (ITU) and the Australian Research Council (ARC). Finally I thank the School of Computer Science and Engineering for providing me with access to a computer cluster to do my simulations.

Abstract

Sparse signal processing has far-reaching applications including compressed sensing, media compression/denoising/deblurring, microarray analysis and medical imaging. The main reason for its popularity is that many signals have a sparse representation given that the basis is suitably selected. However the difficulty lies in developing an efficient method of recovering such a representation.

To this aim, two efficient sparse signal recovery algorithms are developed in the first part of this thesis. The first method is based on direct minimization of the l_0 norm via cyclic descent, which is called the L0LS-CD (l_0 penalized least squares via cyclic descent) algorithm. The other method minimizes smooth approximations of sparsity measures including those of the l_0 norm via the majorization minimization (MM) technique, which is called the QC (quadratic concave) algorithm.

The L0LS-CD algorithm is developed further by extending it to its multivariate (V-L0LS-CD (vector L0LS-CD)) and group (gL0LS-CD (group L0LS-CD)) regression variants. Computational speed-ups to the basic cyclic descent algorithm are discussed and a greedy version of L0LS-CD is developed. Stability of these algorithms is analyzed and the impact of the penalty parameter and proper initialization on the algorithm performance are highlighted. A suitable method for performance comparison of sparse approximating algorithms in the presence of noise is established. Simulations compare L0LS-CD and V-L0LS-CD with a range of alternatives on under-determined as well as over-determined systems.

The QC algorithm is applicable to a class of penalties that are neither convex nor concave but have what we call the quadratic concave property. Convergence proofs of this algorithm are presented and it is compared with the Newton algorithm, concave convex (CC) procedure, as well as with the class of proximity algorithms. Simulations focus on the smooth approximations of the l_0 norm and compare them with other l_0 denoising algorithms.

Next, two applications of sparse modeling are considered. In the first application the L0LS-CD algorithm is extended to recover a sparse transfer function in the presence of coloured noise. The second uses gL0LS-CD to recover the topology of a sparsely connected network of dynamic systems. Both applications use Laguerre basis functions for model expansion.

The role of model selection in sparse signal processing is widely neglected in literature. The tuning/penalty parameter of a sparse approximating problem should be selected using a model selection criterion which minimizes a desired discrepancy measure. Compared to the commonly used model selection methods, the SURE (Stein's unbiased risk estimator) estimator stands out as one which does not suffer from the limitations of other methods. Most model selection criterion are developed based on signal or prediction mean squared error. The last section of this thesis develops an SURE criterion instead for parameter mean square error and applies this result to l_1 penalized least squares problem with grouped variables. Simulations based on topology identification of a sparse network are presented to illustrate and compare with alternative model selection criteria.

Contents

List of Figures

List of Tables

List of Acronyms

AIC	Akaike's Information Criterion
AP	Alternating Projections
BELM	Backward Elimination
BIC	Bayesian Information Criterion
CC	Concave Convex
CIC	Covariance Inflation Criterion
cIHT	Constrained Iterative Hard Thresholding
CoSaMP	Compressive Sampling Matching Pursuit
DC	Difference of Convex
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transforms
DWT	Discrete Wavelet Transform
FIR	Finite Impulse Response
FN	False Negatives
FOCUSS	Focal Under-determined System Solver
FP	False Positives
FPR	False Positive Rate
FSEL	Forward Selection
gL0LS	Group l_0 Penalized Least Squares criterion
gL0LS-CD	Group l_0 Penalized Least Squares via Cyclic Descent
JPEG	Joint Photographic Experts Group
IALZ	Iterative Approximate l_0 norm
IHT	Iterative Hard Thresholding
JLZA	Joint l_0 Approximation
L0LS	l_0 Penalized Least Squares criterion

L0LS-CD	l_0 Penalized Least Squares via Cyclic Descent
L1LS	l_1 Penalized Least Squares criterion
L1LS-CD	l_1 Penalized Least Squares via Cyclic Descent
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
M-FOCUSS	FOCUSS for Multiple Measurement Vectors
MAP	Maximum A-Posterior
MDL	Minimum Description Length
MM	Majorization Minimization
MMV	Multiple Measurement Vectors
MOD	Method of Optimal Directions
MP	Matching Pursuit
MSE$_\mu$	Signal Mean Squared Error
MSE$_\beta$	Parameter Mean Squared Error
NP	Nondeterministic Polynomial time
OLS	Orthogonal Least Squares
OMP	Orthogonal Matching Pursuit
PCA	Principal Component Analysis
pIHT	Penalized Iterative Hard Thresholding
QC	Quadratic Concave
QC-G$_{0,\gamma}$	QC algorithm with the G_0^γ penalty
QC-TH$_{0,\gamma}$	QC algorithm with the $\mathrm{TH}_0^{\gamma,b}$ penalty
ReMBo	Reduce MMV and Boost
RIC	Risk Inflation Criterion
SCAD	Smoothly Clipped Absolute Deviation
SCSI	Sparse Coloured System Identification
SCSI-L0	l_0 penalized SCSI
SCSI-L1	l_1 penalized SCSI
SNR	Signal to Noise Ratio
SOMP	Simultaneous Orthogonal Matching Pursuit
StOMP	Stagewise Orthogonal Matching Pursuit
SURE	Stein's Unbiased Risk Estimator
SV	Singular Value
SVD	Singular Value Decomposition

TN	True Negatives
TP	True Positives
TPR	True Positive Rate
V-L0LS	Vector l_0 Penalized Least Squares criterion
V-L0LS-CD	Vector l_0 Penalized Least Squares via Cyclic Descent
V-L1LS	Vector l_1 Penalized Least Squares criterion
V-L1LS-CD	Vector l_1 Penalized Least Squares via Cyclic Descent

List of Notations

I_m	Identity matrix of dimension m
$X_{n\times p}$	Matrix X has n rows and p columns
$x_{(j)}$	j$^{\text{th}}$ column of X matrix
$x_{[j]}$	j$^{\text{th}}$ row of X matrix
x_{rc}	Element at the r$^{\text{th}}$ row and c$^{\text{th}}$ column of the X matrix
$\beta_{n\times 1}$	β vector has n elements
$\hat{\beta}$	Estimate of the β vector
β^k	Estimate of the β vector at the k$^{\text{th}}$ iterate of an iterative algorithm
$\beta^\star$	True or exact value of the β vector
β_j	j$^{\text{th}}$ element of the β vector
$\lvert x\rvert$	Absolute value of x
$\max(a, b)$	Maximum of the two values a or b, i.e. if $a < b$, then $\max(a, b) = b$
$\min(a, b)$	Minimum of the two values a or b, i.e. if $a < b$, then $\min(a, b) = a$
sgn(x)	Sign of x, sgn(x)=1 if $x > 0$, and -1 if $x < 0$
X^T	Transpose of the X matrix
X^{-1}	Inverse of the X matrix
trace(X)	Trace of the X matrix
diag(β)	Matrix with the elements of β at the diagonal
$I(\cdot)$	The indicator function. It equals 1 if the condition holds true and 0 otherwise. e.g. $I(\beta_j \neq 0) = 1$ if $\beta_j \neq 0$ and is 0 if $\beta_j = 0$.
$\lvert\lvert\beta\rvert\rvert_0$	l_0 norm of the β vector, $\lvert\lvert\beta\rvert\rvert_0 = \sum I(\beta_j \neq 0)$
$\lvert\lvert\beta\rvert\rvert_q$	l_q norm of the β vector, $\lvert\lvert\beta\rvert\rvert_q = \sum \left(\sum \lvert\beta_j\rvert^q\right)^{1/q}$
$\lvert\lvert\beta\rvert\rvert_1$	l_1 norm of the β vector, $\lvert\lvert\beta\rvert\rvert_1 = \sum \lvert\beta_j\rvert$
$\lvert\lvert\beta\rvert\rvert$	Euclidean norm of the β vector, $\lvert\lvert\beta\rvert\rvert = \sqrt{\beta^T\beta}$
$\lvert\lvert\beta\rvert\rvert_\Gamma$	Weighted Euclidean norm of the β vector, $\lvert\lvert\beta\rvert\rvert_\Gamma = \sqrt{\beta^T\Gamma\beta}$

$\|\|B\|\|_{r,0}$	Vector l_0 norm of the B matrix, $\|\|B\|\|_{r,0} = \sum I(\|\|\beta_{[j]}\|\|_r \neq 0)$
$\|\|B\|\|_{r,q}$	Vector l_q norm of the B matrix, $\|\|B\|\|_{r,q} = \sum \left(\sum \|\|\beta_{[j]}\|\|_r^q\right)^{1/q}$
$\|\|B\|\|_{r,1}$	Vector l_1 norm of the B matrix, $\|\|B\|\|_{r,1} = \sum \|\|\beta_{[j]}\|\|_r$
$\|\|\|\beta\|\|\|_{r,0}$	Group l_0 norm of the β vector, if the coefficients of β vector are grouped as $\beta = [\bar{\beta}_1^T, \cdots \bar{\beta}_m^T]^T$, then $\|\|\beta\|\|_0 = \sum I(\|\|\bar{\beta}_j\|\|_r \neq 0)$
$\|\|\|\beta\|\|\|_{r,q}$	Group l_q norm of the β vector, if the coefficients of β vector are grouped as $\beta = [\bar{\beta}_1^T, \cdots \bar{\beta}_m^T]^T$, then $\|\|\beta\|\|_q = \sum \left(\sum \|\|\bar{\beta}_j\|\|_r^q\right)^{1/q}$
$\|\|\|\beta\|\|\|_{r,1}$	Group l_1 norm of the β vector, if the coefficients of β vector are grouped as $\beta = [\bar{\beta}_1^T, \cdots \bar{\beta}_m^T]^T$, then $\|\|\beta\|\|_1 = \sum \|\|\bar{\beta}_j\|\|_r$
G_0^γ	Smooth approximation of the l_0 norm, $\mathrm{G}_0^\gamma = \sum \left(1 - e^{-\beta_j^2/2\gamma^2}\right)$
$\mathrm{TH}_0^{\gamma,b}$	Smooth approximation of the l_0 norm, $\mathrm{TH}_0^{\gamma,b} = \sum \tanh\left(\frac{\|\beta_j\|^b}{\gamma}\right)$
$f(\beta)$	Function of the β variable
$f'(\beta)$	First derivative of the function $f(\cdot)$ with respect to the β variable
$f''(\beta)$	Second derivative of the function $f(\cdot)$ with respect to the β variable
dim(A)	Dimension of the set A, i.e number of elements of the set
h	Penalty parameter
$R_{\beta,h}$	Parameter mean squared error
$\hat{R}_{\beta,h}$	Estimate of the parameter mean squared error
$R_{\mu,h}$	Signal mean squared error
$\hat{R}_{\mu,h}$	Estimate of the signal mean squared error
h_b	Value of h that minimizes $R_{\beta,h}$
h_m	Value of h that minimizes $R_{\mu,h}$
h_β	Value of h that minimizes $\hat{R}_{\beta,h}$
h_μ	Value of h that minimizes $\hat{R}_{\mu,h}$
h_{BIC}	Value of h that minimizes the BIC criterion

Chapter 1

Introduction

Sparse signal approximation aims at using a minimum number of elementary signals from a dictionary to find a good approximation of a signal of interest. Sparse regression is widely used in many engineering, statistics, and applied mathematics applications. Section 1.1 discusses some applications which motivated the development of sparse signal processing. Then sections 1.2 to 1.5 defines sparsity, discusses measures of sparsity and presents an overview of commonly used algorithms for sparse modeling.

Recently, multivariate and group regression variants of sparse signal estimation problems have gained a lot of interest. Sections 1.6 and 1.7 gives an introduction to them and presents the motivation for their development.

Any regularization method requires the selection of a regularization penalty parameter. Many model selection criteria have been developed to select these parameters by comparing competing models using various discrepancy measures. Section 1.8 gives an overview of the commonly used model selection criteria along with the advantages and disadvantages of each method. Finally section 1.9 gives and outline of this thesis and section 1.10 gives the list of its contributions.

1.1 Sparse Modeling

The Nyquist-Shannon sampling theorem states that a bandlimited analog signal can be exactly represented from its samples if it is sampled uniformly at a rate

at least twice as fast as the signal's highest frequency. This theory has enabled analog signals to be processed using digital signal processing tools. However in many applications sampling at the Nyquist rate results in a high volume of data, which makes it difficult to process, transmit or store signals. This generates the need for a more efficient method of signal representation.

It has been found that many media (images [142, 276], audio [185], video [272]) signals can be approximated by a sparse representation with respect to suitably selected bases. Widely used media encoding standards such as JPEG [184] and JPEG-2000 [232] exploit this fact to compress images. Both standards represent the images using a new coordinate system (base) which results in sparse coordinates which are then processed to encode the image. JPEG relies on the discrete cosine transform (DCT) [229] while JPEG-2000 relies on the discrete wavelet transform (DWT) [134]. Other image processing problems such as image denoising [41, 59] and image deblurring [46, 60] also relies on sparse representations of images.

In applications such as medical imaging [129, 139] and radar imaging [186, 13] signal acquisition can be dangerous, expensive or difficult. This has motivated compressed sensing [47, 27] which uses fewer measurements than what is demanded by traditional sampling theory. Provided that the signal is compressible, i.e. it has a sparse representation with respect to a known base, compressed sensing provides a way to recover the signal using few linear functionals. Although compressed sensing enables the recovering of signals using fewer measurements, it comes at a price. When traditional sampling theory is used, signals can be recovered by applying simple linear reconstruction formulas. However, the task of recovering a signal from reduced measurements require nonlinear, relatively expensive reconstruction techniques.

The concept of sparse signal approximation is widely used in many more applications such as oceanic engineering [132], antennas and propagation [1], machine learning [94], support vector machines [17], blind source separation [133, 81], modeling of natural languages [84, 210], face recognition [262], microarrays [182] etc.

The task of finding a sparse model of a signal can be cast as a problem of finding the sparse solution of a system of linear equations. Consider the following

linear regression model,

$$y_{n\times 1} = s_{n\times 1} + \varepsilon = X_{n\times p}\beta_{p\times 1} + \varepsilon = \sum_{j=1}^{p} x_{(j)}\beta_j + \varepsilon. \tag{1.1}$$

In applications such as media compression, y is the media signal being compressed and in applications such as signal denoising and medical imaging, y is the noisy data or measurement vector. s is the noise free signal and X is the regression matrix, dictionary or base with respect to which we seek a sparse approximation of s. Throughout this thesis references have to be made to both rows and columns of matrices such as X. So the following compact notation is used,

$$X = [x_{rc}] = [x_{(1)}, \ldots, x_{(d)}] = \begin{bmatrix} x_{[1]}^T \\ \vdots \\ x_{[p]}^T \end{bmatrix}.$$

The columns of X are often referred to as predictors or atoms. In many cases it is assumed that X is column scaled, so that $\|x_{(j)}\| = 1$; it is well known that this improves the numerical conditioning of the X matrix [14]. β is the vector of regression coefficients or parameters. ε is the noise which corrupted the original signal s. In a sparse approximation problem y and X are known and the objective is to estimate β or to recover the noise free signal s. The parameter estimate is denoted as $\hat{\beta}$ and the signal estimate is denoted as $\hat{s} = X\hat{\beta}$.

The term sparse representation is often used in noise free systems and when $y = \hat{s}$ and the term sparse approximation is used in noisy systems thus $y \simeq \hat{s}$. Two error terms can be defined for the latter; signal estimation error $(s - \hat{s})$ and residual $(y - \hat{s})$. Signal estimation error cannot be directly calculated since s is unknown.

Since in practice almost all problems are noise affected this thesis is focussed on sparse approximation of noisy signals.

1.2 Dictionaries

The properties of the dictionary vary greatly depending on the application. In regression problems the number of samples (n) is greater than the number of predictors (p), thus the resulting systems are over-determined. In inverse problems however the dictionaries are over-complete, thus the number of samples is less than the number of predictors, and the resulting systems are under-determined.

Using a dictionary which comprises of the minimum number of basis vectors is usually only adequate to sparsely represent a small class of signals. Thus over-complete dictionaries are formed using a carefully chosen set of redundant basis vectors such that one general dictionary can represent a larger class of signals. Popular over-complete dictionaries are steerable wavelets, segmented wavelets, Gabor dictionaries, multiscale Gabor dictionaries, curvelets, contourlet, wedgelet, bandlet etc. [59, 36].

The choice of the dictionary that sparsifies the signals is crucial for the success of sparse representation [194]. In general, the choice of a proper dictionary can be done either by building a dictionary based on a mathematical model of the data or by selecting a dictionary from a set of candidates using dictionary learning techniques [141, 239, 173].

[172] provided a key contribution to the area of dictionary learning by training a dictionary for sparse representation of small image patches collected from a number of natural images. This inspired a series of subsequent works which were mostly focused on statistical training methods. These methods were either based on Maximum Likelihood estimation [130] or Maximum A-Posterior (MAP) estimation [124].

Popular dictionary learning techniques are MOD (method of optimal directions) [65, 66], union of orthobases [128], generalized PCA (principal component analysis) [258] and the K-SVD algorithm [3, 59]. Theoretical guarantees of the uniqueness of over-complete dictionaries are given in [2]. Uniqueness depends on the quantity and nature of the data set and the sparsity of the desired representation.

Although over-complete dictionaries have the advantage of being able to represent a wide variety of signals, the solution to an under-determined system of

equations is not unique. So additional constraints are needed to recover the best suited representation of the signal s. As mentioned above many applications require a sparse representation of s.

1.3 Goals of Sparse Modeling

Definition I: Sparsity - A vector or a matrix is called sparse when most of its coefficients are zero.

Thus finding a sparse representation of s with respect to X is equivalent to representing s as a linear combination of few columns of X i.e. $\hat{\beta}$ has few non-zero coefficients. Given y and X the goals of a sparse modeling problem can be formally presented as follows,

1. **Sparsity** - The solution ($\hat{\beta}$) to equation (1.1) should have fewer non-zero coefficients than the min(n, p), where n is the length of y and p is the number of predictors.

2. **Reconstruction Error**

    - **Signal Reconstruction Error**- Linear combination of the selected atoms should provide the best approximation of the signal s.
    - **Parameter Reconstruction Error**- Estimate of the coefficient vector ($\hat{\beta}$) should resemble the original coefficient vector (β) as closely as possible.

These two objectives may contradict each other. Signal reconstruction error can generally be reduced when more atoms are used for the approximation but this reduces the sparsity of the solution. Thus some tradeoff has to be made, and it is usually determined by the regularization tuning parameter.

1.4 Measures of Sparsity

Although the qualitative definition of sparsity seem simple and straightforward, a universally accepted quantitative measurement of this concept does not exist.

A sparse measure (diversity measure) is a mapping of a vector or a matrix to a real number such that its value decreases as the sparsity increase i.e. number of non-zero coefficients decrease. Many measures of sparsity have been used in the literature and there is a debate on the properties that they should have. Consider a sparse vector β of length p with elements $\beta_j, j = 1, \cdots, p$. Fourteen commonly used measures of sparsity are listed below,

$$l_0 \quad : \quad \|\beta\|_0 = \sharp\{j, \beta_j \neq 0\} = \sum I(\beta_j \neq 0), \tag{1.2a}$$

$$l_0^\gamma \quad : \quad \|\beta\|_{0,\gamma} = \sharp\{j, |\beta_j| > \gamma\}, \tag{1.2b}$$

$$l_q \quad : \quad \|\beta\|_q = \left(\sum |\beta_j|^q\right)^{1/q}, 0 < q < 1 \tag{1.2c}$$

$$l_1 \quad : \quad \|\beta\|_1 = \sum |\beta_j|, \tag{1.2d}$$

$$\mathrm{TH}_0^{\gamma,b} \quad : \quad \sum \tanh\left(\frac{|\beta_j|^b}{\gamma}\right), \gamma > 0, b > 0, \qquad (l_0 \text{ approximation}) \tag{1.2e}$$

$$\mathrm{G}_0^\gamma \quad : \quad \sum \left(1 - e^{-\beta_j^2/2\gamma^2}\right), \qquad (l_0 \text{ approximation}) \tag{1.2f}$$

$$\mathrm{sqrt}_1^\gamma \quad : \quad \sum \sqrt{\beta_j^2 + \gamma^2} - \gamma, \qquad (l_1 \text{ approximation}) \tag{1.2g}$$

$$\mathrm{LC}_1^\gamma \quad : \quad \gamma \sum \ln\cosh\left(\frac{\beta_j}{\gamma}\right), \qquad (l_1 \text{ approximation}) \tag{1.2h}$$

$$\log \quad : \quad \sum \log\left(|\beta_j| + \gamma\right) - \log\left(\gamma\right), \tag{1.2i}$$

$$\text{Kurtosis} \quad : \quad \kappa_4 = \frac{\sum \beta_j^4}{\left(\sum \beta_j^2\right)^2} \tag{1.2j}$$

$$\text{Gini} \quad : \quad 1 - 2\sum \frac{\beta_{\{j\}}}{\|\beta\|_1}\left(\frac{p - j + \frac{1}{2}}{p}\right) \tag{1.2k}$$

$$H_G \quad : \quad \sum \ln \beta_j^2 \qquad \text{(Gaussian entropy)} \tag{1.2l}$$

$$H_S \quad : \quad \sum \bar{\beta}_j \ln \bar{\beta}_j, \text{ where } \bar{\beta}_j = \frac{\beta_j^2}{\|\beta\|^2} \qquad \text{(Shannon entropy)} \tag{1.2m}$$

$$u_\theta \quad : \quad \min_{i=1,\cdots,p-\lceil p\theta\rceil+1} \frac{\beta_{\{i+\lceil p\theta\rceil-1\}} - \beta_{\{i\}}}{\beta_{\{p\}} - \beta_{\{1\}}} \tag{1.2n}$$

$$\text{s.t. } \lceil p\theta\rceil \neq p \tag{1.2o}$$

l_0 simply counts the number of non-zero coefficients and $I(\cdot)$ is the indicator function. Sparsity measures u_θ and Gini are calculated after rearranging the

data in the ascending order such that, $\beta_{\{1\}} \leq \beta_{\{2\}} \leq \cdots \leq \beta_{\{p\}}$. The Gini Index was originally proposed in economics as measure of the inequality of wealth [136, 43, 85]. The kurtosis measures the peakedness of a distribution and u_θ measures the smallest range which contains a certain percentage of the data. All except seven sparsity measures (l_0, l_1, H_G, H_S, κ_4, u_θ and Gini) depend on tuning parameters.

$\mathrm{TH}_0^{\gamma,b}$ and G_0^{γ} are smoothed approximations of the l_0 norm. To see this, consider firstly $\mathrm{TH}_0^{\gamma,b}$ with b=1 (this argument is valid as long as $b > 0$),

$$\begin{aligned} \beta_j = 0 \quad &: \quad \tanh\left(\frac{|\beta_j|}{\gamma}\right) = 0, \\ \beta_j \neq 0 \quad &: \quad \lim_{\gamma\to 0} \tanh\left(\frac{|\beta_j|}{\gamma}\right) = \lim_{\gamma\to 0} \frac{1 - e^{-\frac{2|\beta_j|}{\gamma}}}{1 + e^{-\frac{2|\beta_j|}{\gamma}}} = 1. \end{aligned}$$

Now consider G_0^{γ},

$$\begin{aligned} \beta_j = 0 \quad &: \quad 1 - e^{-\beta_j^2/2\gamma^2} = 0, \\ \beta_j \neq 0 \quad &: \quad \lim_{\gamma\to 0} 1 - e^{-\beta_j^2/2\gamma^2} = 1. \end{aligned}$$

Similarly sqrt_1^{γ} and LC_1^{γ} are smoothed approximations of the l_1 norm. First consider sqrt_1^{γ},

$$\begin{aligned} \beta_j = 0 \quad &: \quad \sqrt{\beta_j^2 + \gamma^2} - \gamma = 0, \\ \beta_j \neq 0 \quad &: \quad \lim_{\gamma\to 0} \sqrt{\beta_j^2 + \gamma^2} - \gamma = |\beta_j|. \end{aligned}$$

Now consider LC_1^{γ},

$$\beta_j = 0 \ : \ \gamma \sum \ln\cosh\left(\frac{\beta_j}{\gamma}\right) = 0,$$

$$\beta_j > 0 \;:\; \lim_{\gamma\to 0} \gamma \ln \cosh\left(\frac{\beta_j}{\gamma}\right) = \lim_{\gamma\to 0} \gamma \ln\left(\frac{e^{\frac{\beta_j}{\gamma}} + e^{-\frac{\beta_j}{\gamma}}}{2}\right),$$

$$= \lim_{\gamma\to 0} \gamma \ln\left(\frac{e^{\frac{\beta_j}{\gamma}}\left(1 + e^{-\frac{2\beta_j}{\gamma}}\right)}{2}\right) = \beta_j + \lim_{\gamma\to 0} \gamma\left(ln\left(1 + e^{-\frac{2\beta_j}{\gamma}}\right) - 1\right) = \beta_j$$

$$\beta_j < 0 \;:\; \lim_{\gamma\to 0} \gamma \ln \cosh\left(-\frac{|\beta_j|}{\gamma}\right) = \lim_{\gamma\to 0} \gamma \ln\left(\frac{e^{-\frac{|\beta_j|}{\gamma}} + e^{\frac{|\beta_j|}{\gamma}}}{2}\right),$$

$$= \lim_{\gamma\to 0} \gamma \ln\left(\frac{e^{\frac{|\beta_j|}{\gamma}}\left(1 + e^{-\frac{2|\beta_j|}{\gamma}}\right)}{2}\right) = |\beta_j| + \lim_{\gamma\to 0} \gamma\left(ln\left(1 + e^{-\frac{2|\beta_j|}{\gamma}}\right) - 1\right) = |\beta_j|$$

Properties of these sparsity measures are discussed in [114, 119, 191, 188]. In a financial setting [43] presents four properties that a measure of inequality of wealth distribution should have. Since measuring the distribution of wealth should have similar properties to measuring the energy distribution of coefficients, these properties were adapted by the signal processing community [191, 114].

1. **Robin Hood** - (Daltons 1st Law) Robin Hood decreases sparsity. Stealing from the rich and giving to the poor, decreases the inequity of wealth distribution (assuming you do not make the rich poor and the poor rich).

2. **Scaling** - (Daltons modified 2nd Law) Sparsity is scale invariant. Multiplying wealth by a constant factor does not alter the effective wealth distribution.

3. **Rising Tide** - (Daltons 3rd Law) Adding a constant decreases sparsity. Give everyone a trillion dollars and the small differences in overall wealth are then negligible.

4. **Cloning** - (Daltons 4th Law) Sparsity is invariant under cloning. If you have a twin population with identical wealth distribution, the sparsity of wealth in one population is the same for the combination of the two.

Two additional favorable properties of a measure of sparsity have been introduced in [191] and developed further in [114].

5. **Bill Gates** - Bill Gates increases sparsity. As one individual becomes infinitely wealthy, the wealth distribution becomes as sparse as possible.

6. **Babies** - Babies increase sparsity. Adding individuals with zero wealth to a population increases the sparseness of the distribution of wealth.

[114] and [191] have both recommended Gini index as it is the only measure that satisfies all six criterion given above. [119] suggests that kurtosis should not be used and recommend $\mathrm{TH}_0^{\gamma,b}$ for measuring sparseness in noisy data.

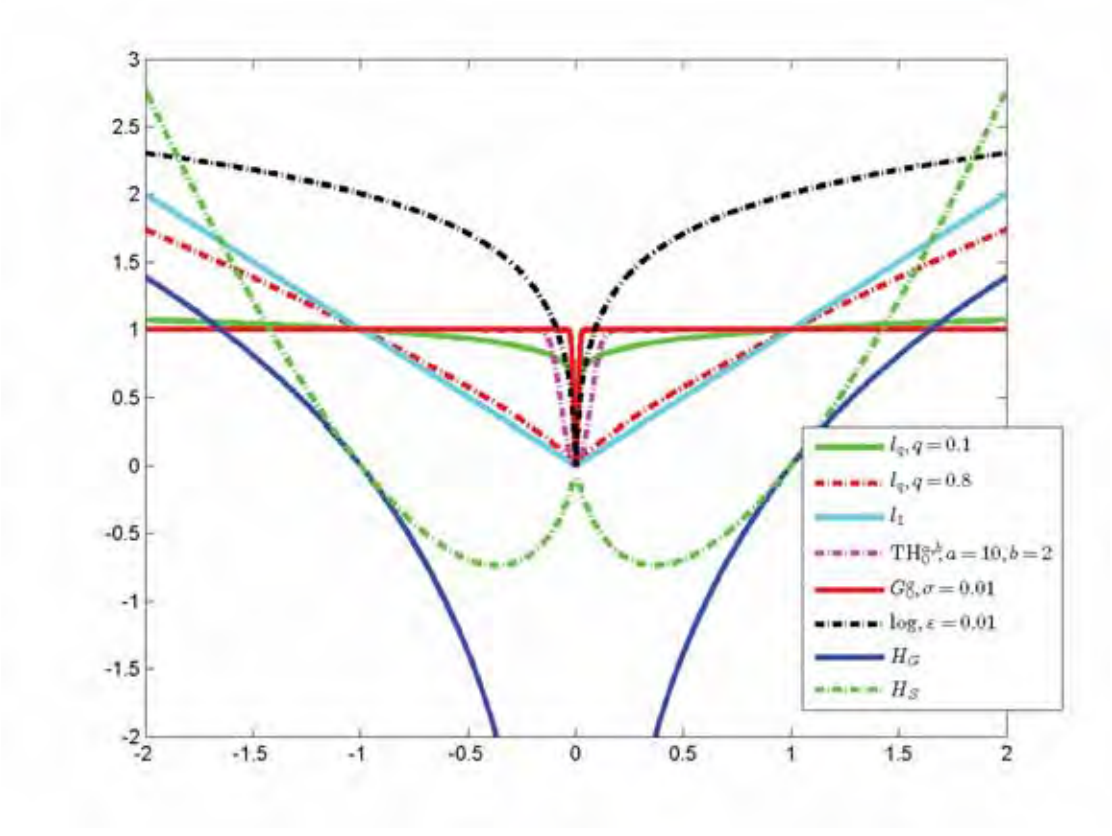

Figure 1.1: Measures of sparsity as a function of component amplitude

The l_0 measure is severely criticized in all three articles for the following reasons,

1. The magnitude of non-zero elements is ignored (violate properties 1 and 5 listed above).

   (a) An infinitesimally small value is treated the same as a large value. Thus the presence of noise makes the l_0 measure completely inappropriate.

   (b) A change in the value of a non-zero element does not change the value of the l_0 norm measure. When a coefficient changes slightly, a corresponding change in the value of the sparsity measure is expected based on the importance of the particular coefficient to the overall sparsity. l_0 does not show this property.

2. The derivative of the measure contains no information and majority of the optimization methods fail when the l_0 norm is used.

The first allegation is ill founded in the sparse regression setting. Generally the importance of a coefficient is not measured by its value but rather by its contribution to the signal approximation, i.e. to minimizing the mean squared error $\|y - X\hat{\beta}\|^2$ or residual. Furthermore when the l_0 norm is used in combination with the least squares criterion the first problem of the above list can be eliminated because the least squares term will generate amplitude dependence. This will be further investigated in section 1.5.

The second point is the major reason for the lack of popularity of the l_0 norm. Due to its discrete non-convex nature, finding a global minimum of a criterion including an l_0 term is NP hard [162, 44, 164]. Thus smooth functions such as $\mathrm{TH}_0^{\gamma,b}$ and G_0^{γ} have been used with the aim of approximating the l_0 norm.

Measures of sparsity as a function of component amplitude are given in figure 1.1. All the measures except for Shannon entropy minimize at zero. Shannon entropy prefer components to be at a non-zero value less than 1. This quality of the Shannon entropy to produce concentrated non sparse solutions has been previously observed by [188]. Gaussian entropy does not have an upper or lower bound. The log function (1.2i) is generally used as an approximation of l_0. However it does not have an upper bound, thus it cannot be a true approximation of the l_0 norm. It is best viewed as a sparsity measure in its own right. l_q with q

close to zero closely approximates the l_0 norm. However l_q is non convex and as q approaches zero the criterion has many local minimum. Thus it becomes difficult to find the global minimum of such a function.

1.5 Sparse Modeling Algorithms

In a noise free system, a sparse solution to equation (1.1) can be obtained by optimizing the following criterion,

$$\min_{\beta} f(\beta) \qquad \text{s.t.} \quad y = X\beta. \tag{1.3}$$

where $f(\beta)$ is one of the sparsity measures given in section 1.4. In a noisy system finding a sparse solution to the linear system of equations is not so straightforward. However an approximate solution can be achieved by incorporating sparsity measures given in section 1.4 with the least squares criterion. Thus the goals presented in section 1.3 can be achieved by optimizing one of the following three criteria,

$$\min_{\beta} \ f(\beta) \qquad \text{s.t.} \quad \|y - X\beta\|^2 \leq \epsilon, \tag{1.4a}$$

$$\min_{\beta} \ \|y - X\beta\|^2 \qquad \text{s.t.} \quad f(\beta) \leq k, \tag{1.4b}$$

$$\min_{\beta} \ \|y - X\beta\|^2 \ + \ h \ f(\beta), \tag{1.4c}$$

where ϵ, k and h are tuning parameters. The least squares term $\|y - X\beta\|^2$ measures the residual and $f(\beta)$ promotes sparsity. The relevant tuning parameter determines the relative emphasis given to the two terms of the criterion. Methods of tuning parameter selection will be discussed in section 1.8. Although the l_0 norm by itself is insensitive to the values of the coefficients, amplitude dependance is provided by the least squares term. Thus as stated in section 1.4 the criticism of the l_0 norm, given in [119, 191, 114] are irrelevant in the sparse regression setting.

Criteria (1.4b) and (1.4c) are the most popular forms used in the literature. (1.4a) and (1.4b) are constrained criteria and (1.4c) is a penalized criterion. In

other words (1.4c) is a qualitative dual of (1.4a) and (1.4b). If $f(\beta)$ is convex then we can establish a quantitative equivalence between the solutions of the different forms of the criterion [24]. However when $f(\beta)$ is non convex establishing a formal quantitative duality is not straightforward and may not hold.

An overview of existing sparsity producing algorithms are presented next.

1.5.1 Exponential and Randomized Search Methods

For a general X matrix, when $f(\beta) = l_0$, the only known method for finding the global minimum of any of the above mentioned criteria (1.4a)-(1.4c) is via exponential search methods such as exhaustive search, branch and bound [163] etc. Exponential search methods have gained popularity in the area of feature selection [95, 45], they have the advantage of having high accuracy but at the cost of high complexity which increases exponentially with the number of predictors. Thus these methods are impractical for large scale problems. The signal processing community has developed some interest in these methods [234, 235] but due to the computational burden they have failed to gain much popularity.

Random search methods such as simulated annealing [122, 26] and genetic algorithms [106, 86, 87] attempt to find optimal solutions by searching in a random fashion. They are designed to escape local minima and their complexity is generally low. The accuracy of these algorithms are high [255, 212] provided that the control parameters are properly selected but the choice of such parameters is not straightforward. Random search methods have been applied on isolated signal processing applications [231] but they have not managed to gain much interest.

1.5.2 Greedy Methods

Greedy pursuits such as forward selection(FSEL), backward elimination (BELM) [53, 57, 155], CLEAN [105, 201] and orthogonal matching pursuit (OMP) [240, 244, 246] are classical methods used to solve linear inverse problems. They iteratively refine the sparse solution by identifying one or more coefficients that can be added, removed or refined to improve the estimation in the least squares sense. Traditionally these algorithms are stopped with an ad-hoc stopping rule.

The CLEAN algorithm [105, 201] is widely used in astronomy and it can be traced back to 1930s [224, 233]. This same algorithm has been rediscovered by the signal processing community and is referred to as matching pursuit (MP) [145]. A modified CLEAN algorithm that can perform l_1 denoising was presented in [220].

OMP is closely related to the CLEAN algorithm. At each iteration the coefficients are selected using the same principles as the CLEAN algorithm but the estimate and the error signal (residual) are calculated in a different manner.

FSEL is also known as orthogonal least squares (OLS) [35]. Unlike CLEAN and OMP, FSEL selects the coefficients that has the largest partial correlation with the residual and like OMP, at each iteration, the estimate β is calculated by orthogonally projecting y on to the predictors of the active set. An overview of FSEL, CLEAN and OMP is given in appendix 2.A.

Instead of adding predictors to the active set, BELM starts by estimating β using all the predictors and then removes ones that have the smallest contribution to reconstruction error at each iteration.

Hybrid methods such as Efroymson's algorithms [155] have been introduced with the hope of capturing the benefits of both FSEL and BELM. Other variants include adding, removing or refining more than one predictor at a time and weighting the predictors, giving larger weights to predictors which are more likely to be part of the active set.

Stagewise orthogonal matching pursuit (StOMP) [52], and compressive sampling matching pursuit (CoSaMP) [166] are improvements of the OMP algorithm. In contrast to OMP, StOMP and CoSaMP allows many coefficients to enter the model at each iteration. Furthermore StOMP terminates after a fixed number of iterations. However as discussed in section 2.11.4 these algorithms can only support a limited sparsity range.

1.5.3 Optimization Methods

These algorithms are based on optimizing a constrained (1.4b) or penalized (1.4c) least squares formulation. The $f(\beta) = l_0$ penalty is the most appropriate as it is a direct measure of sparsity thus it promotes maximum sparsity. However as

mentioned in section 1.4 the discrete, non-convex nature of the criterion poses the greatest difficulty in finding its global minimum. Thus other measures of sparsity have gained more popularity.

1.5.3.1 l_1 Norm

One can approximate the l_0 norm by a convex penalty, with the l_1 norm (1.2d) being the most common. Due to its convexity there is a direct relationship between the penalized and constrained versions of the criterion [24, 175, 267]. The l_1 norm is used by many algorithms such as [6, 236, 36, 56, 74, 278]. The most famous of which are the least absolute shrinkage and selection operator (LASSO) [236] and least angle regression (LARS) algorithm [56]. These have gained popularity because (1.4b) or (1.4c) with $f(\beta) = l_1$ norm can be easily solved using either a quadratic programming approach [236, 249, 267], homotopy approaches [175], coordinate wise optimization [73] or gradient projection method [70]. A variant of these methods is the non-negative garrotte [25] which starts with the least squares estimate and then scales it by a vector obtained by solving an l_1 constrained least squares criterion.

Many workers have researched the performance of sparse approximation with the l_1 norm. Initial studies on noiseless systems showed that under certain conditions the solution to (1.3) with $f(\beta) = l_0$ can also be obtained by using the l_1 norm [50, 49]. Similarly [48] presents conditions for noisy systems where the solution to (1.4a) with $f(\beta) = l_0$ can be obtained by using the l_1 convex relaxation. [281, 267, 242] gives criteria under which l_1 penalized or constrained least squares solution selects the true model, i.e. the support set of the estimate is the same as that of the original coefficient vector.

However [68] have shown that the l_1 norm penalty produces biased estimates since its derivative does not vanish for large values of the coefficients. Therefore coefficients with large values are favored by the l_1 norm penalty. Variations of the l_1 norm such as the clipped l_1 penalty and smoothly clipped absolute deviation (SCAD) are introduced to overcome this drawback [68, 8, 67]. Furthermore the conditions under which the l_1 norm perform well are somewhat limited and it has been shown in many instances that its solutions are of low sparsity and

therefore undesirable [76, 189, 190]. Simulation results given in section 2.11.5 further confirms this.

1.5.3.2 l_q Norm

Incorporation of the l_q norm into regression was first explored by [126, 72]. Bridge regression involves optimizing (1.4c) with $f(\beta) = \sum |\beta_j|^q$ for $q \geq 0$ [72]. This encompasses ridge regression ($q = 2$), which was introduced by [104, 103], the l_1 penalty, which was discussed above and the l_0 penalty as special cases. Although [72] pointed out that it is desirable to optimize with respect to q they did not provide an algorithm to solve bridge regression for any value of q.

Ridge regression ($q = 2$) produces a solution to an ill-conditioned inverse problem with reduced variance compared to that of an ordinary least squares solution, however its solutions are non-sparse. [74] has presented an algorithm for optimizing bridge regression for $q \geq 1$ where the tuning parameter and q are chosen using generalized cross-validation. It has been shown that the performance can be improved by variably selecting q in the open range $(\infty, 1]$ rather than keeping it fixed at $q = 2$ or $q = 1$.

In bridge regression sparsity of the estimate is promoted when $0 < q \leq 1$. However the resulting optimization problem is no longer convex when $q < 1$. The main interest of this penalty is that it is a quasi smooth approximation of the l_0 norm when q tends to zero. A simplex search algorithm was presented to solve (1.3) with $f(\beta) = l_q$ norm in [126]. In a series of papers [91, 90, 188] focal under-determined system solver (FOCUSS[1]) algorithm was developed which uses a re-weighed norm minimization technique to apply the l_q penalty on noiseless systems. The regularized FOCUSS algorithm [187] extended this concept to noisy systems. An algorithm based on majorization minimization technique for solving (1.4c) with $f(\beta) = \sum |\beta|^q, 0 < q < 1$ is given in [150].

The performance of the l_q norm on noiseless systems was studied in [31, 32, 71]. They have shown that in compressed sensing fewer measurements are needed to reconstruct a sparse signal when the l_q norm is used than would be expected with the l_1 norm. They have also derived the conditions under which the performance

[1]The p tuning parameter of FOCUSS is referred to as q in this thesis.

of the l_q norm can be guaranteed and these conditions are less restrictive than that of the l_1 norm. The performance of the l_q norm on noisy systems was studied in [108, 195]. [108] derived the conditions under which the l_q norm penalized least squares solution recovers the true model and [195] have shown that the l_q norm with $q < 1$ provides better theoretical guarantees in terms of stability and robustness than the l_1 norm.

1.5.3.3 Smoothed Approximations of l_0 Norm

Alternatively the l_0 norm can be approximated by a differentiable function. The G_0^γ function (1.2f) is used to approximate the l_0 norm in [158, 159] where a gradient based method is used to find the minimum. This idea is further developed in [115, 117] which uses the same approximation but presents a better structured algorithm to find the solution (IALZ). However all these algorithms handle noise in an ad-hoc manner. [116] presents a vector version (JLZA) of [117], and the derivation given there can support noise. The log penalty (1.2i) and $TH_0^{\gamma,b}$ function (1.2e) are other common smoothed approximations of the l_0 norm, although, as indicated in section 1.4, the log penalty is better regarded as a sparsity measure in its own right. The log penalty is discussed in [269, 28] and the $TH_0^{\gamma,b}$ function is used in [16].

1.5.3.4 Direct Iterative Minimization

The other approach is direct iterative minimization of (1.4b) or (1.4c) with $f(\beta) = l_0$. An iterative procedure based on alternating projections (AP) is presented in [147] and it concentrates on solving the l_0 constrained least squares criterion. Landweber based iterative hard thresholding (IHT) algorithm have been proposed in [100] and later developed by [20].

Two different IHT algorithms have been presented in [20]. One algorithm minimizes the l_0 penalized least squares criterion (pIHT) while the M-sparse-IHT algorithm minimizes an l_0 constrained least squares criterion (cIHT). pIHT is presented as an algorithm to refine the solution found with methods such as matching pursuit or basis pursuit denoising. The authors of [20] have dismissed pIHT as being much inferior to the cIHT algorithm. Section 2.11.5 shows how

they can be made comparable. The cIHT algorithm was further developed in [19] and a method of increasing its speed is given in [18].

1.5.4 Bayesian Methods

In Bayesian terms, most of the algorithms based on sparsity measures discussed above can be viewed as performing standard MAP estimation using a fixed, sparsity-inducing prior. The sparse Bayesian learning algorithms [271, 238] use a parameterized prior and learn the prior as opposed to MAP methods that use a fixed prior. Bayesian methods are not used in this thesis and will not be discussed any further.

1.6 Multivariate Regression

Multivariate sparse regression is also known as sparse representation of multiple measurement vectors or as the joint sparse recovery problem. Multivariate over-determined regression has a long history in statistics [202, 7] while the under-determined case has had much less attention particularly in a sparse setting. The sparse version has been motivated by applications such as neuromagnetic inverse problems [89], direction-of-arrival estimation [116], channel equalization [69], and array processing [118] where the multivariate regression problem naturally arises. Although the computational burden of multivariate regression is higher than the scaler version, it has been shown that the prediction accuracy of the estimates can be greatly improved by simultaneously optimizing multiple measurement vectors.

A number of sparse scalar regression algorithms have been extended to the multivariate case. Simultaneous orthogonal matching pursuit (SOMP) is presented in [245, 93]. Algorithms that minimize multivariate versions of the penalized or constrained least squares criterion were also developed. [241] and [144] presents algorithms based on the vector l_1 norm and [40] presents that of the vector l_q norm which is called the regularized M-FOCUSS algorithm. The vector l_0 norm is approximated by a vector G_0^{γ} function in [116] which develops the JLZA algorithm. The ReMBo algorithm [157] converts the multivariate regression to a scalar regression by randomly combining the measurement vectors. The perfor-

mance of multivariate regression algorithms have been compared under various conditions [34, 62, 257] but most of the work has been done on noiseless systems.

1.7 Regression of Grouped Variables

Many problems of economics [165], neuroscience [23, 37] and biology [261, 75] can be cast as a network topology identification problem which requires sparse regression of grouped variables. Application to topology identification of a sparsely connected network will be addressed in chapter 5. Other applications which require regression of grouped variables include sparse channel estimation [274], cognitive spectrum sensing [42], colour imaging [143] etc.

Therefore many scalar regression algorithms have been extended to handle group sparsity. Group sparse versions of greedy algorithms include extensions of matching pursuit [274, 63], clustered orthogonal matching pursuit [197] and cycling orthogonal least squares [151]. Groups sparse versions of l_1 and l_q norm optimization methods have also been developed. Group LASSO [278] is a very common algorithm used in regression of grouped variables. Group versions of LARS and non-negative garrotte have also been developed in [278]. Other group sparse algorithms that optimize the l_1 norm is given in [64, 63] and extension of the FOCUSS algorithm is given in [143].

1.8 Model Selection

As shown in section 1.5 the optimization criterion (1.4a)-(1.4c) of a regularization problem require selection of a penalty or tuning parameter. The least squares term of these criteria promote prediction accuracy and $f(\beta)$ promotes sparsity. The tuning parameter determines the emphasis given to the two terms, thus it determines the sparsity of the estimate. Therefore the model that is fitted to the observed data will then depend on the value of the tuning parameter.

The performance of competing models can be assessed using a discrepancy measure of some kind. The definition of the discrepancy will vary depending on the requirements of the application. Model selection criteria for selecting the

tuning parameters are typically based on minimizing an estimate of the desired discrepancy [135].

Tuning parameters can be discrete e.g. model dimension, iteration count or continuous e.g. threshold level. Model selection criteria can be broadly divided into two categories; nonparametric methods and model based methods also known as covariance penalties [135]. Only some of the model selection criteria can handle both discrete and continuous tuning parameters.

1.8.1 Nonparametric Methods

Cross-validation, bootstrap techniques and L-curve are nonparametric methods. In cross-validation [161, 228] the data are subdivided into two parts, one part is used for estimation and the other part is used for validation. This method has been widely applied, especially to linear problems. However for nonlinear problems cross-validation is computationally intensive because it involves solving the whole inverse problem for all possible divisions of the data set. Generalized cross-validation which is a rotation-invariant version of ordinary cross-validation is presented in [263, 88]. Bootstrap methods [54] generate an estimator of the expected discrepancy by resampling. However [38] have shown that the bootstrap criterion has a downward bias and [55] have shown that model based methods perform better than cross-validation and bootstrap methods. The L-curve method originated in [156] and was later developed by [97, 98]. The L-curve method has gained popularity partly because it is computationally cheap. However it has been heavily criticized in [96, 260].

1.8.2 Model Based Methods

Mallows's C_p [146], Akaike's information criterion (AIC) [4], the Bayesian information criterion (BIC) [200] (which is generally the same as the minimum description length (MDL) principle [192]) and Stein's unbiased risk estimator (SURE) [226] are model based selection criteria. Unlike the nonparametric methods C_p, AIC and BIC can usually only handle discrete tuning parameters. In the applications given in this thesis we are able to overcome that problem. These methods are computationally cheap but are biased because they assume that the

data are generated from a model in the model class being fitted. This is rarely the case in practice. More recently generalizations and modifications of these criterion have been developed which include risk inflation criterion (RIC) [79], covariance inflation criterion (CIC) [237] and Cauchy prior modification of BIC [120].

Mallows's C_p, AIC, BIC, RIC and CIC select the best model for the linear regression model (1.1) contaminated by Gaussian noise with known variance ($\varepsilon \sim N(0, \sigma^2 I)$) by optimizing a model selection criterion of the form,

$$\|y - \hat{\mu}\|^2 + \lambda d\sigma^2. \tag{1.5}$$

where $\hat{\mu} = X\hat{\beta}$ and d is the number of non-zeros in $\hat{\beta}$. The first term promotes the goodness of fit and the second term penalizes the complexity of the fitted model while $\lambda > 0$ determines the trade off between the two terms. Mallows's C_p and AIC set $\lambda = 2$, BIC set $\lambda = \log(n)$, RIC set $\lambda = 2\log(p)$ and CIC set $\lambda = 4\sum_{j=1}^{d} \log(n/j)/d$. However when the penalty parameter λ is fixed the resulting criterion is effective and consistent only under specific conditions thus limiting its applicability [204, 282, 209]. For instance when λ is large the criterion is likely to perform well when the size of the true model is small or the true model is sparse. Thus attempts have been made to adaptively select the penalty parameter according to the application [80, 207, 206], however as shown in [277] it is not clear whether this is a successful method. [205] combines a class of modeling procedures into a unified framework with the hope of combining the strengths of different modeling procedures.

1.8.3 SURE

SURE is a model based selection criterion and it does not suffer from the limitations of the model selection criterion discussed so far. Unlike the other model based methods it can handle continuous as well as discrete tuning parameters and although it makes some assumptions SURE does not assume that fitted model is the same as the operating model. SURE is computationally cheap and it does not require iteration. It is an exact method in that it does not use any approximations such as Taylor series expansion or linearization.

SURE was introduced by [226] and was first used as a tuning parameter selection method by [51, 110]. Subsequently [213] suggested that SURE could have wide applicability and it was applied to a range of ill-conditioned inverse problems such as optical flow [167, 168, 208], nonparametric signal estimation [218], anisotropic diffusion [216, 217], total variation denoising [215], rank selection PCA [251, 250] support vector machines [219] and image processing [138, 149]. Other work includes [61]. All the existing literature on SURE is based on the signal (prediction) mean squared error. Chapter 6 develops a model selection criterion based on parameter mean squared error via SURE.

1.9 Thesis Outline

Chapter 2 is based on l_0 denoising. The first half of this chapter investigates the ability of existing algorithms to perform l_0 denoising. FSEL, CLEAN and OMP are possible candidates for l_0 denoising because these algorithms reduce the l_0 penalized least squares (L0LS) criterion at each iteration provided a proper stopping rule is selected. However this chapter shows that their estimates are not guaranteed to satisfy the optimality conditions of a local minimum of the L0LS criterion. Despite indications in the noise free case to the contrary in [187], there seem to be a belief in the literature that regularized FOCUSS with $q = 0$ does l_0 denoising. This chapter shows that this is also false.

In the second half of this chapter a novel algorithm based on cyclic descent to minimize the L0LS criterion is developed, which is called L0LS-CD (l_0 penalized least squares via cyclic descent). Then it is extended to the multivariate (V-L0LS-CD) and group (gL0LS-CD) regression variants. Computational speed ups are discussed and a greedy version of L0LS-CD is developed. Stability of these algorithms are analyzed and simulations compare L0LS-CD and V-L0LS-CD with alternatives. and The following publications relate to this chapter.

A. J. Seneviratne and V. Solo, "On Exact L0 Denoising," *Submitted to IEEE Transactions on Signal Processing.*

A. J. Seneviratne and V. Solo, "On Vector L0 Penalized Multivariate Regression," in *Proc. IEEE International Conference on Acoustics, Speech*

and Signal Processing, Kyoto, Japan, 2012, pp. 3613-3616.

B. Cassidy, V. Solo and A. J. Seneviratne, "Grouped L0 Least Squares Penalised Magnetoencephalography," in *Proc. IEEE International Symposium on Biomedical Imaging*, Barcelona, Spain, 2012, pp. 868-871.

Chapter 3 develops an algorithm based on the majorization minimization technique which can optimize the least squares criterion penalized with any penalty obeying the quadratic concave property defined in chapter 3. We call it the QC (quadratic concave) algorithm. Convergence analysis of the QC algorithm is provided and it is compared with the Newton algorithm, concave convex (CC) procedure, as well as with the class of proximity algorithms. The material is closely related to the following publication.

V. Solo and A. J. Seneviratne, "The Quadratic Concave Algorithm," *in preparation for submission to IEEE Transactions on Signal Processing.*

Chapter 4 is based on an application of sparse modeling. A method for sparse transfer function estimation in the presence of coloured noise is developed. Unlike most previous sparse transfer function estimation methods, this approach via Laguerre polynomials guarantees stability of the fitted transfer functions. Also unlike previous methods the new procedure can handle coloured noise. Both l_1 and l_0 penalized procedures are discussed. The material covered in this chapter is based on the following publication.

A. J. Seneviratne and V. Solo, "Sparse Coloured System Identification with Guaranteed Stability," in *Proc. IEEE Conference on Decision and Control*, Honolulu, Hawaii, 2012, pp. 2826-2831.

Chapter 5 is also based on an application of sparse modeling. Here the problem of identifying the topology of a sparsely connected network of dynamic systems is addressed. The goal is to identify the links, the direction of information flow and the transfer function of each dynamic system. The output of each system is affected by the incoming data of the directly connected systems and noise. In contrast to the related existing work, causal Laguerre basis functions are used to

expand the transfer functions and l_0 penalty is used to enforce sparsity. Since the network is sparsely connected the system topology is estimated using gL0LS-CD. This chapter is based on the following publication.

> A. J. Seneviratne and V. Solo, "Topology Identification of a Sparse Dynamic Network," in *Proc. IEEE Conference on Decision and Control*, Honolulu, Hawaii, 2012, pp. 1518-1523.

Chapter 6 develops a model selection criterion. Any regularization method requires the selection of a regularization penalty parameter. Many model selection criteria have been developed to compare competing models using various discrepancy measures. Most model selection methods are focused on signal mean squared error i.e. prediction mean squared error. This chapter develops a model selection criterion based on parameter mean squared error via SURE. Then it is applied to group LASSO. Simulation results based on topology identification of a sparse network are presented to illustrate and compare with alternative model selection criteria. This chapter is based on the following publication.

> A. J. Seneviratne and V. Solo, "Parameter Based Model Selection for the Group LASSO via SURE," *Submitted to IEEE Transactions on Signal Processing.*

1.10 Contributions

1. Chapter 2: l_0 denoising.

    - L0LS-CD algorithm - minimizes the l_0 penalized least squares criterion via cyclic descent.
    - Multivariate (V-L0LS-CD) and group (gL0LS-CD) regression variants of L0LS-CD.
    - Convergence analysis of the L0LS-CD algorithm.

- Discussion of computational speed-ups to the basic cyclic descent algorithm and the development of a greedy version of L0LS-CD.
- Investigate the ability of existing algorithms to perform l_0 denoising.
- Establish a suitable method for performance comparison of sparse approximating algorithms in the presence of noise.
- Comparison of the performance of the L0LS-CD and V-L0LS-CD algorithms with a range of sparse algorithms, on under-determined and over-determined systems, highlighting the importance of proper tuning parameter selection and initialization.

2. Chapter 3: QC algorithm.

   - QC algorithm - minimizes the least squares criterion penalized with any quadratic concave penalty via majorization minimization technique.
   - Convergence analysis of the QC algorithm.
   - Comparison of the QC algorithm with the Newton algorithm, concave convex (CC) procedure, as well as with the class of proximity algorithms.
   - Comparison of the performance of the QC algorithm with respect to two penalties that approximate the l_0 penalty (G_0^{γ}, $\mathrm{TH}_0^{\gamma,b}$).

3. Chapter 4: Application of sparse modeling - transfer function estimation.

   - SCSI algorithm - method for sparse transfer function estimation on systems with coloured noise via cyclic descent.
   - This method guarantees the stability of the estimated model by the use of Laguerre basis functions.
   - Comparison of the l_0 norm and l_1 norm variants of the SCSI algorithm.

4. Chapter 5: Application of sparse modeling - network topology identification.

- Development of a method of topology identification of a sparsely connected network via gL0LS-CD with Laguerre basis functions for model expansion.
- Comparison of gL0LS-CD and group LASSO algorithms based on this application.

5. Chapter 6: Parameter Based Model Selection via SURE.

- Development of a general SURE criterion for parameter mean square error.
- This result is applied to obtain an SURE criterion for parameter mean square error of the l_1 penalized least squares problem with grouped variables.
- Comparison of this criterion with other model selection criteria by a simulation based on topology identification of a sparse network.

Chapter 2

l_0 Denoising

The importance and motivation behind sparse modeling was given in section 1.1. Many sparse modeling problems can be posed as finding a sparse solution to a linear regression model (1.1). This thesis is based on finding sparse solutions to systems with noise and the goals of sparse approximation are given in section 1.3. These goals can be achieved by optimizing one of the criterion (1.4a)-(1.4c) with an appropriate sparsity measure. An overview of commonly used sparsity measures was given in section 1.4.

The l_0 norm has been widely criticized in literature, mainly for not being dependent on the amplitude of the coefficients. As discussed in section 1.4 these allegation are not relevant in the sparse regression setting as the least squares term in the optimization criterion provides the amplitude dependance. Thus in sparse regression the l_0 penalty is the most favorable as it is a direct measure of sparsity and therefore would promote more sparsity.

Optimizing one of the criterion (1.4a)-(1.4c) with $f(\beta) = l_0$ norm is called l_0 denoising. However finding a global minimum to such a criterion is known to be NP hard, thus an exhaustive search is the only guaranteed method known so far. This is the main reason for the unpopularity of the l_0 norm.

This chapter is based on l_0 penalized least squares (L0LS) criterion and its multivariate and grouped variable variants. The scalar L0LS criterion is developed in section 2.1 and the conditions for its local minimum are given in section 2.1.1. Sections 2.2 to 2.4 investigates the ability of existing algorithms to perform l_0 denoising. Section 2.5 develops a cyclic descent based algorithm to optimize

the L0LS criterion which we call L0LS-CD (l_0 least squares via cyclic descent[1]) followed by termination criterion and computational speed-ups. Two different variation of the L0LS-CD algorithm is also developed in this chapter. Vector l_0 penalized least squares (V-L0LS) criterion is developed in 2.6 and variant of L0LS-CD that can handle multivariate regression (V-L0LS-CD[2]) is developed in section 2.7. Similarly l_0 penalized least squares criterion for grouped variables (gL0LS) is given in section 2.8 and section 2.9 develops a variant of the L0LS-CD algorithm to handle grouped variables (gL0LS-CD). Non-trivial stability analysis is developed in section 2.10. Simulation results are presented in sections 2.11 and 2.12. Conclusions are in section 2.13.

2.1 l_0 Penalized Least Squares (L0LS)

The aim is to find a sparse solution to the linear regression problem (1.1) by optimizing the scale free l_0 penalized least squares criterion,

$$\frac{\|y - X\beta\|^2}{\sigma^2} + h_0\|\beta\|_0, \tag{2.1}$$

where σ^2 is the noise variance. This chapter assumes that the X matrix is column scaled, so that $\|x_{(u)}\| = 1$. As mentioned in chapter 1, this will improve the numerical conditioning of the X matrix. If we multiply across by σ^2 we can replace h_0 by $h = h_0\sigma^2$. This removes σ^2 but makes it clear that h is then scale dependent. We then obtain (1.4c) with $f(\beta) = l_0$ norm,

$$J(\beta) = \|y - X\beta\|^2 + h\|\beta\|_0. \tag{2.2}$$

In this thesis minimizing $J(\beta)$ is referred to as the L0LS problem.

[1] A cyclic descent l_0 algorithm has been described in [153] but it differs from ours as discussed below.

[2] [203] was based on this algorithm.

2.1.1 Conditions for a local minimum

Since $J(\beta)$ is not even differentiable it is not immediately obvious how to find conditions for a local minimum. For the l_1 penalized least squares problem this was done by [6] (which predates [236]). In a very neat piece of analysis, building on work of [243], the conditions for, local minimum of $J(\beta)$ were given by [20], as follows,

Theorem I: Optimality Conditions [20].

Define the index sets $\Gamma_0 = \{j : \dot{\beta}_j = 0\}$, $\Gamma_c = \{j : \dot{\beta}_j \neq 0\}$ and set $\dot{\gamma}_j = x_{(j)}^T(y - X\dot{\beta})$. Then $\dot{\beta}$ is a local minimum of $J(\beta)$ iff,

(Ia) $|\dot{\gamma}_j| \leq \sqrt{h}$, $j \in \Gamma_0$.

(Ib) $\dot{\gamma}_j = 0$, $j \in \Gamma_c$.

(Ic) $|\dot{\beta}_j| \geq \sqrt{h}$, $j \in \Gamma_c$.

Note that from (Ib), $\dot{\beta}$ is just the least squares estimate of β using only the indices specified in Γ_c.

2.2 pIHT

The pIHT algorithm developed in [20] can find a local minimum of $J(\beta)$, thus does l_0 denoising. Landweber iteration based pIHT calculates new values for all the coefficients at each iteration. The coefficient vector is updated only after all p new coefficients are calculated.

$$\beta^k = H_{\sqrt{h}}(\beta^{k-1} + X^T(y - X\beta^{k-1})),$$

where k is the iteration counter and $H_{\sqrt{h}}$ is the element wise hard thresholding operator,

$$H_{\sqrt{h}}(\beta_u) = \beta_u I(|\beta_u| > \sqrt{h}),$$

where β_u is the u^{th} coefficient of the β vector. pIHT is discussed further below.

2.3 Regularized FOCUSS with q=0

Despite indications in the noise free case to the contrary in [187], there seems to be a belief that regularized FOCUSS with $q = 0$ does l_0 denoising. The following proof shows that it does not solve the L0LS problem; rather it solves an entropy penalized least squares problem.

We need only to provide a counter example and to do this the simple case where $X = I$ is sufficient. The regularized FOCUSS algorithm then reduces to,

$$\beta_u^k = \frac{|\beta_u^{k-1}|^{2-q}}{|\beta_u^{k-1}|^{2-q} + h} \; y_u, \quad 1 \leq u \leq p.$$

We deduce $\text{sgn}(\beta_u^k) = \text{sgn}(y_u)$ and so the iteration becomes,

$$\begin{aligned} |\beta_u^k| &= \frac{|\beta_u^{k-1}|^{2-q} \, |y_u|}{|\beta_u^{k-1}|^{2-q} + h}, \\ &= |\beta_u^{k-1}| \; \frac{|y_u|}{D_u^{k-1}}. \end{aligned} \tag{2.3}$$

where $D_u^{k-1} = |\beta_u^{k-1}| + h/|\beta_u^{k-1}|^{1-q}$. For $q = 0$ we have,

$$D_u^{k-1} = |\beta_u^{k-1}| + \frac{h}{|\beta_u^{k-1}|},$$

and it is easily seen that this has a minimum at $\sqrt{h}$ of value $2\sqrt{h}$. Thus $D_u^{k-1} \geq 2\sqrt{h}$. Thus,

$$\begin{aligned} |\beta_u^k| &\leq |\beta_u^{k-1}| \; \frac{|y_u|}{2\sqrt{h}}, \\ &\leq \left(\frac{|y_u|}{2\sqrt{h}}\right)^{k-1} |\beta_u^0|. \end{aligned}$$

so if $|y_u| < 2\sqrt{h}$ we find $|\beta_u^{k-1}| \to 0$. On the other hand from (2.3), $|\beta_u^k| < |y_u|$ so $|\beta_u^{k-1}|$ is a bounded sequence and so must have at least one limit point.

From (2.3), if the sequence converges it will be to a fixed point which obeys

(for $q = 0$),

$$|\beta_u| = \frac{|\beta_u|^2 |y_u|}{|\beta_u|^2 + h},$$
$$\Rightarrow |\beta_u|^2 + h = |\beta_u||y_u| \tag{2.4}$$

This quadratic equation has solutions,

$$|\beta_u| = \frac{|y_u| \pm \sqrt{y_u^2 - 4h}}{2} \triangleq \Psi_{\pm}(|y_u|)$$

So there are only 2 possible limit points. However we already saw $|\beta_u^{k-1}| < |y_u|$ so both limit points can be reached.

So the solution is $I(|y_u| \geq 2\sqrt{h})\Psi(|y_u|)$. However the L0LS problem $\min_\beta \|y - \beta\|^2 + h\sum_1^p I(\beta_u \neq 0)$ has solution $y_u I(|y_u| > \sqrt{h})$, which is clearly not the same. So regularized FOCUSS with $q = 0$ does not converge to the solution of the L0LS problem. In fact for $|\beta| > 0$ it is easily seen that (2.4) is the optimality condition for the minimizer of $\frac{1}{2}\|y - \beta\|^2 + h\sum_1^p \log|\beta_u|$.

This observation may be regarded as an extension to the noisy case, of the connections made in [188] with Gaussian entropy ($\sum_1^p \log|\beta_u|$) in the noise free case.

2.4 Other Candidate Algorithms

FSEL, CLEAN and OMP are widely used algorithms in sparse signal approximation. They are initialized with $\beta^0 = 0$ and at each iteration they select a new index to be included in the active set. They differ in the method by which they select the new index and by the way they update the estimate and the residual. For the sake of completeness the outline of these algorithms are presented in section 2.A.

At the k^{th} iteration, given the estimate β^k and residual e^k, let X^k be the k columns of X that are already selected. After another iteration we have β^{k+1} and e^{k+1}. The newly selected column added to the active set at the $k + 1^{th}$ iteration

is given respectively by,

$$\begin{aligned} &\text{CLEAN, OMP:} && x_{(\hat{u})} : \hat{u} = \arg.\max_{u \notin \Gamma_c^k} |\gamma_u^k| \\ &\text{FSEL:} && x_{(\hat{u})} : \hat{u} = \arg.\max_{u \notin \Gamma_c^k} \frac{|\gamma_u^k|}{\Delta_k} \end{aligned}$$

where $\gamma_u^k = x_{(u)}^T e^k$, $\Gamma_c^k = \{j : \beta_j^k \neq 0\}$ and where

$$\begin{aligned} &\text{FSEL, OMP:} && \Delta_k^2 = 1 - \rho^T (X^{k,T} X^k)^{-1} \rho, \\ & && \rho = x_{(\hat{u})}^T X^k. \\ &\text{CLEAN:} && \Delta_k = 1. \end{aligned}$$

The update of the energy of the error signal for FSEL, CLEAN and OMP is given by,

$$\|e^{k+1}\|^2 = \|e^k\|^2 - \frac{(x_{(\hat{u})}^T e^k)^2}{\Delta_k^2} \tag{2.5}$$

Now we compute the change in J(β). We have,

$$J(\beta^{k+1}) = J^{k+1} = \|e^{k+1}\|^2 + h \sum I(\beta_u^{k+1} \neq 0)$$

Since one new component is added to the active set at each iteration in all three algorithms, we find, via (2.5)

$$\begin{aligned} J^{k+1} - J^k &= \|e^{k+1}\|^2 - \|e^k\|^2 + hI(\beta_{\hat{u}}^{k+1} \neq 0) \\ &= -\frac{(x_{(\hat{u})}^T e^k)^2}{\Delta_k^2} + h \end{aligned}$$

Thus $J(\beta)$ is reduced while $(x_{(\hat{u})}^T e^k)^2 / \Delta_k^2 > h$. Therefore provided we stop when $|x_{(\hat{u})}^T e^k| \leq \sqrt{h}\Delta_k$, FSEL, CLEAN and OMP algorithms reduce $J(\beta)$ at each iteration. Thus these algorithms are possible candidates for l_0 denoising under this stopping rule.

Such a stopping rule does not correspond to the ad-hoc conditions used to terminate FSEL, CLEAN and OMP [246]. One might hope then that this new

stopping rule would allow these algorithms to converge to a local minimum of $J(\beta)$.

Here a simple example is given to show that OMP estimates do not satisfy the L0LS optimality conditions given in section 2.1.1. Consider the following X matrix and $\beta^\star$ vector,

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0.5 & 0 & 0.7071 \\ 0 & 0 & 0 & 1 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0.7071 & 0.7071 & 0 \end{bmatrix}$$

$$\beta^\star = \begin{bmatrix} 0 & 0 & 0.6211 & 0 & 0 & 0.7015 & 0 \end{bmatrix}^T$$

Assume the noise is Gaussian with zero mean and variance $\sigma^2 = 0.0878$. From (1.1), $y = [0.1887 \quad -0.3646 \quad 0.6496 \quad -0.3241 \quad 0.4610]^T$. Set $h = 0.125$, then OMP produces the following estimate when stopping rule $|x_{(\hat{u})}^T e^k| \leq \sqrt{h}\Delta_k$ is used,

$$\hat{\beta} = \begin{bmatrix} 0 & -0.6906 & 0.3236 & 0 & 0.6520 & 0 & 0 \end{bmatrix}^T$$

However $\sqrt{h} = 0.3536$. Thus (Ic) of the optimality conditions is violated. Therefore OMP do not do l_0 denoising. Similar examples can be given for FSEL and CLEAN.

2.5 L0LS-CD

Cyclic descent has been used to solve the l_1 penalized least squares in [74, 278, 73, 183]. Here cyclic descent is applied to (2.2). While the idea of using cyclic descent to minimize (2.2) seems to have been in the folklore, we know of no published work.

The basic concept of this iterative procedure is to fix all the coefficients of β at their current value except for one and minimize the criterion with respect to the selected coefficient. Once the new value of the selected coefficient is calculated, β

is updated immediately and then the criterion is minimized with respect to the next coefficient.

This is the major difference between cyclic descent and Landweber iteration based algorithms. Cyclic descent is like a classic Gauss-Seidel algorithm where new coefficient values are used immediately to update the next coefficient. Landweber is like a classic Jacobi algorithm where new coefficient values are not used to update the estimate until the new values of all the coefficients are calculated. Because of this L0LS-CD is expected to be faster than pIHT.

From (2.2), elementary algebra gives, for any pair β, β^o

$$\begin{aligned} J(\beta) =& J(\beta^o) - 2(\beta - \beta^o)^T\gamma^o + (\beta - \beta^o)^T X^T X(\beta - \beta^o) \\ &+ h\sum_1^p I(\beta_j \neq 0) - h\sum_1^p I(\beta_j^o \neq 0). \end{aligned} \tag{2.6}$$

where $\gamma^o = X^T(y - X\beta^o)$. Given the iterate k, factor $k = lp + u$ where l is an integer and $1 \leq u \leq p$. Thus at iterate $k-1$, coefficient $u-1$ was updated and the next iterate will update the coefficient with the index u.

Set $\gamma_u^{k-1} = x_{(u)}^T(y - X\beta^{k-1})$. From (2.6) the change in the value of the criterion at consecutive iterations is given by,

$$\begin{aligned} \Delta J_k =& J(\beta^k) - J(\beta^{k-1}) \\ =& -2\Delta_u^k\gamma_u^{k-1} + (\Delta_u^k)^2 + hI(\beta_u^k \neq 0) - hI(\beta_u^{k-1} \neq 0), \end{aligned} \tag{2.7}$$

where $\Delta_u^k = \beta_u^k - \beta_u^{k-1}$. Denote $J_u^{k-1} = I(\beta_u^{k-1} \neq 0)$.

To minimize $J(\beta^k)$ two cases must be considered.

If $\beta_u^k = 0$,

$$\begin{aligned} \Delta J_k &= 2\beta_u^{k-1}\gamma_u^{k-1} + (\beta_u^{k-1})^2 - hJ_u^{k-1}, \\ &= (\beta_u^{k-1} + \gamma_u^{k-1})^2 - (\gamma_u^{k-1})^2 - hJ_u^{k-1}. \end{aligned} \tag{2.8}$$

If $\beta_u^k \neq 0$, differentiation leads to the minimum being achieved at $\Delta_u^k = \gamma_u^{k-1}$.

So that,

$$\beta_u^k = \beta_u^{k-1} + \gamma_u^{k-1}, \tag{2.9}$$

and

$$\Delta J_k = -(\gamma_u^{k-1})^2 + h - hJ_u^{k-1}. \tag{2.10}$$

When comparing the two cases, it is clear that $\beta_u^k = 0$ is the minimizer if,

$$\begin{aligned}(\beta_u^{k-1} + \gamma_u^{k-1})^2 - (\gamma_u^{k-1})^2 - hJ_u^{k-1} &< -(\gamma_u^{k-1})^2 + h - hJ_u^{k-1} \\ \equiv \quad (\beta_u^{k-1} + \gamma_u^{k-1})^2 &< h \\ \equiv \quad |\beta_u^{k-1} + \gamma_u^{k-1}| &< \sqrt{h}\end{aligned}$$

We additionally have to make a choice as to how to define β_u^k when $|\beta_u^{k-1}+\gamma_u^{k-1}| = \sqrt{h}$. Because then both $\beta_u^k = 0$ and $\beta_u^k = \beta_u^{k-1} + \gamma_u^{k-1}$ deliver the same value for ΔJ_k. Here the choice is made as follows.

When $|\beta_u^{k-1} + \gamma_u^{k-1}| = \sqrt{h}$ then,
if $\beta_u^{k-1} \neq 0$ set $\beta_u^k = \beta_u^{k-1} + \gamma_u^{k-1}$,
if $\beta_u^{k-1} = 0$ set $\beta_u^k = 0$.

We shall see that these choices deliver lemma D given in section 2.10. To sum up we now have the following update[1].

If $\beta_u^{k-1} \neq 0$ set

$$\beta_u^k = (\beta_u^{k-1} + \gamma_u^{k-1}) I(|\beta_u^{k-1} + \gamma_u^{k-1}| \geq \sqrt{h}). \tag{2.11}$$

If $\beta_u^{k-1} = 0$ set

$$\beta_u^k = \gamma_u^{k-1} I(|\gamma_u^{k-1}| > \sqrt{h}). \tag{2.12}$$

For future reference (2.11) and (2.12) are written as,

$$\beta_u^k = A(\beta_u^{k-1}, \gamma_u^{k-1}). \tag{2.13}$$

[1]Note that this update differs from that in [153]

Also in the sequel, notation $I^k_{u,<} = I(|\beta^{k-1}_u + \gamma^{k-1}_u| < \sqrt{h})$ etc. will be used for simplicity. From (2.8), (2.10) and (2.13) the change in the criterion is then given as follows.

If $\beta^{k-1}_u \neq 0$

$$\begin{aligned}\Delta J_k =&[(\beta^{k-1}_u + \gamma^{k-1}_u)^2 - (\gamma^{k-1}_u)^2 - h]I^k_{u,<} - (\gamma^{k-1}_u)^2 I^k_{u,\geq},\\ =&[(\beta^{k-1}_u + \gamma^{k-1}_u)^2 - h]I^k_{u,<} - (\gamma^{k-1}_u)^2.\end{aligned} \tag{2.14}$$

If $\beta^{k-1}_u = 0$

$$\begin{aligned}\Delta J_k =&[(\beta^{k-1}_u + \gamma^{k-1}_u)^2 - (\gamma^{k-1}_u)^2]I^k_{u,\leq} + [h - (\gamma^{k-1}_u)^2]I^k_{u,>},\\ =&[h - (\gamma^{k-1}_u)^2]I^k_{u,>}.\end{aligned} \tag{2.15}$$

$|\beta^{k-1}_u + \gamma^{k-1}_u|$ and $\sqrt{h}$ are real numbers, therefore the event $|\beta^{k-1}_u + \gamma^{k-1}_u| = \sqrt{h}$ happens with zero probability. Therefore for practical purposes the L0LS-CD update (2.13) is equivalent to,

$$\beta^k_u = (\beta^{k-1}_u + \gamma^{k-1}_u)I(|\beta^{k-1}_u + \gamma^{k-1}_u| \geq \sqrt{h}). \tag{2.16}$$

The L0LS- CD algorithm can be summarized as follows.

L0LS-CD algorithm:

(i) Select β^0 and set iteration counter $k = 1$.

(ii) Decompose $k = lp + u$, where l is an integer and $1 \leq u \leq p$.

(iii) Calculate $\gamma^{k-1}_u = x^T_{(u)}(y - X\beta^{k-1})$.

(iv) Update β_u from (2.16).

(v) Increment k by one and repeat from step (ii) until the termination criterion is met.

The above derivation assumes that $\|x_{(j)}\| = 1, j = 1, ..., p$. By following the same argument, the L0LS-CD update when $\|x_{(j)}\| \neq 1, j = 1, ..., p$ is,

$$\beta^k_u = (\beta^{k-1}_u + \alpha^{k-1}_u)I(\|x_{(u)}\||\beta^{k-1}_u + \alpha^{k-1}_u| \geq \sqrt{h}), \tag{2.17}$$

where $\alpha_u^{k-1} = \gamma_u^{k-1}/\|x_{(u)}\|^2$. Criterion $J(\beta)$ has multiple local minimum thus proper initialization is crucial for the L0LS-CD algorithm. Initialization of the algorithm will be discussed in section 2.11.3. It can be terminated when $J(\beta^k) - J(\beta^{k+1}) \leq$ tolerance or when the algorithm reaches a local minimum such that both conditions (Ia) and (Ib) of section 2.1.1 are met.

2.5.1 Speed Ups

Since the β vector is updated each time a new coefficient is calculated the number of computations needed to calculate γ_u^k can grow extremely high. Here two different updates are presented to reduce the amount of required computations.

(i) For under-determined systems the method introduced in [73] is most effective. Expand γ_u^k as follows,

$$\begin{aligned} \gamma_u^k &= (y - X\beta^k)^T x_{(u)}, \\ &= y^T x_{(u)} - \sum_{j \in \Gamma_c^k} x_{(j)}^T x_{(u)} \beta_j^k, \end{aligned} \tag{2.18}$$

where $\Gamma_c^k = \{j : \beta_j^k \neq 0\}$. Thus γ_u^k can be computed as a linear combination of inner products $y^T x_{(u)}$ and $x_{(j)}^T x_{(u)}$. Initially compute the inner products of each column of X with the y vector. Then each time an index j enters the Γ_c set, the inner product of $x_{(j)}$ with all the columns of X need to be computed.

(ii) For over-determined systems, denote the residual at the k^{th} iteration as $e^k = y - X\beta^k$. Initialize with $e^0 = y - X\beta^0$ and then update according to,

$$e^k = e^{k-1} + (\beta_u^{k-1} - \beta_u^k) x_{(u)}, \tag{2.19}$$

then $\gamma_u^k = x_{(u)}^T e^k$.

2.5.2 Greedy L0LS-CD

An interesting variation of the *shooting* algorithm (which is just L1LS-CD) [74] is presented in [183] called *active-shooting*. Although L0LS-CD converges much

faster than pIHT, its speed can be further improved by incorporating the idea of *active-shooting*. However it was observed that if the idea presented in [183] is applied directly to L0LS-CD it actually increases the execution time of the algorithm. Thus this idea was modified as follows.

This method involves keeping track of the active set of the L0LS-CD algorithm at each iteration. This process is continued until the active set of two consecutive iterations become identical. After that instead of sequentially updating all the coefficients only the coefficients in the active set are updated until the local minimum condition (Ib) is met. Finally the complete β vector has to be updated to check if the active set changes. If the active set remains the same the algorithm can be terminated and if not the process has to be repeated from the beginning.

Greedy L0LS-CD algorithm:

(i) Run the L0LS-CD algorithm and at each iteration update the active set $\Gamma_c^k = \{j : \beta_j^k \neq 0\}$.

(ii) When $\Gamma_c^k = \Gamma_c^{k+1}$, from the next iteration onwards update only the coefficients in the active set until the local minimum condition (Ib) is met.

(iii) Update the complete β vector, $j = 1, ..., p$ according to (2.16). If the active set changes go back to step i. If not return the current β as the final estimate.

The reduction in the computational speed that can be obtained by this greedy version increases as the sparsity increases. This is expected because, as the size of the active set approaches the number of coefficients in the β vector, the effectiveness of the greedy L0LS-CD algorithm reduces.

2.6 Vector l_0 Penalized Least Squares (V-L0LS)

Scalar regression discussed above involve a single measurement vector and produces a single sparse coefficient vector. In contrast multivariate regression considers multiple measurement vectors simultaneously and generates coefficient vectors with a shared sparsity profile. An introduction to multivariate regression and an overview of existing algorithms was given in section 1.6.

Linear regression model (1.1) can be extended to the multivariate case as follows,

$$y_{(c)} = X\beta_{(c)} + \varepsilon, \quad c = 1, \ldots, d.$$

Similar to the scalar regression model, $y_{(c)}$ is a n dimensional measurement vector, $X_{n\times p}$ is a regression matrix or dictionary and $\beta_{(c)}$ is a p dimensional coefficient vector. When d measurement vectors are collected together we can rewrite this as,

$$Y_{n\times d} = X_{n\times p}B_{p\times d} + E, \tag{2.20}$$

where $Y_{n\times d} = [y_{(1)}, \ldots, y_{(d)}]$, $B_{p\times d} = [\beta_{(1)}, \ldots, \beta_{(d)}]$ and $d < n$. This is a multivariate regression model. Since both rows and columns of B needs to be referred in the sequel the following compact notation is used,

$$B = [\beta_{rc}] = [\beta_{(1)}, \ldots, \beta_{(d)}] = \begin{bmatrix} \beta_{[1]}^T \\ \vdots \\ \beta_{[p]}^T \end{bmatrix},$$

and similarly for Y and X matrices. Since the coefficient vectors have the same sparsity profile the resulting B matrix is row sparse. To promote row sparsity, the measures of sparsity discussed in section 1.4 has to be extended to the multivariate setting as follows:

$$\text{Vector } l_0 \quad : \quad \|B\|_{r,0} = \sharp\{j,\ \|\beta_{[j]}\|_r \neq 0\} = \sum I(\|\beta_{[j]}\|_r \neq 0), \tag{2.21a}$$

$$\text{Vector } l_q \quad : \quad \|B\|_{r,q} = \left(\sum \|\beta_{[j]}\|_r^q\right)^{1/q}, 0 < q < 1, \tag{2.21b}$$

$$\text{Vector } l_1 \quad : \quad \|B\|_{r,1} = \sum \|\beta_{[j]}\|_r, \tag{2.21c}$$

$$\text{Vector } \mathrm{G}_0^\gamma \quad : \quad \sum \left(1 - e^{-\|\beta_{[j]}\|_r^2/2\gamma^2}\right), \tag{2.21d}$$

where r is a positive value and $\beta_{[j]}$, a d dimensional vector, is the j^{th} row of the B matrix. Generally $r > 1$ since the objective is to recover a row sparse matrix and sparsity is not enforced within a non-sparse row. Vector l_1 sparsity

measure (2.21c) is used in [241, 144, 257]. [241] set r to ∞, [144] use $r = 2$ and [257] compare the properties of (2.21c) with $r = 1$ and $r = 2$. Vector l_q sparsity measure (2.21b) with $r = 2$ is used in [40] and vector G_0^γ sparsity measure (2.21d) with $r = 2$ is used in [116]. Next section develops an algorithm to minimize the following vector l_0 penalized least squares (V-L0LS) criterion.

$$J(B) = \sum_{c=1}^{d} \|y_{(c)} - X\beta_{(c)}\|^2 + h\|B\|_{r,0}. \tag{2.22}$$

This thesis considers $\|B\|_{r,0}$ with $r = 2$. The optimality conditions of $J(B)$ can be derived in a similar manner to that of the L0LS criterion developed in section 2.1.1 and is given in [203].

The vector l_0 criterion was used previously in another context in [253]; it removes complete rows of B in one go. It should not be confused with the scalar l_0 penalty $\sum_1^p \sum_1^d I(\beta_{rc} \neq 0)$ which only removes individual elements of B.

2.7 V-L0LS-CD

The L0LS-CD algorithm introduced in section 2.5 can be easily extended to minimize the V-L0LS criterion. Similar to equation (2.6), from (2.22), elementary algebra gives, for any pair B, B^o,

$$\begin{aligned} J(B) =& J(B^o) + \sum_{c=1}^{d} \left[-2(\beta_{(c)} - \beta^o_{(c)})^T \gamma^o_{(c)} + (\beta_{(c)} - \beta^o_{(c)})^T X^T X (\beta_{(c)} - \beta^o_{(c)})\right] \\ &+ h \sum_{1}^{p} I(\|\beta_{[j]}\| \neq 0) - h \sum_{1}^{p} I(\|\beta^o_{[j]}\| \neq 0). \end{aligned} \tag{2.23}$$

where $\gamma^o_{(c)} = X^T(y_{(c)} - X\beta^o_{(c)})$. Similar to the L0LS-CD algorithm, given the iterate k, factor $k = lp + u$ where l is an integer and $1 \leq u \leq p$. Set $\gamma^{k-1}_{uc} = x^T_{(u)}(y_{(c)} - X\beta^{k-1}_{(c)})$. Unlike the L0LS-CD algorithm, which updates a single coefficient at an iteration, the V-L0LS-CD algorithm updates an entire row of the B matrix in an iteration. Thus from (2.23) the change in the value of the criterion

at consecutive iterations is given by,

$$\begin{aligned}\Delta J_k =& J(B^k) - J(B^{k-1}) \\ =& \sum_{c=1}^{d} \left[-2\Delta_{uc}^k \gamma_{uc}^{k-1} + (\Delta_{uc}^k)^2\right] + hI(\|\beta_{[u]}^k\| \neq 0) - hI(\|\beta_{[u]}^{k-1}\| \neq 0), \\ =& -2\Delta_{[u]}^k \gamma_{[u]}^{k-1,T} + \|\Delta_{[u]}^k\|^2 + hI(\|\beta_{[u]}^k\| \neq 0) - hI(\|\beta_{[u]}^{k-1}\| \neq 0), \end{aligned} \tag{2.24}$$

where $\Delta_{uc}^k = \beta_{uc}^k - \beta_{uc}^{k-1}$, $\Delta_{[u]}^k = \beta_{[u]}^k - \beta_{[u]}^{k-1}$ and $\gamma_{[u]}^{k-1} = x_{(u)}^T(Y - XB^{k-1})$. Repeating the argument given in section 2.5, the V-L0LS-CD update becomes,

If $\|\beta_{[u]}^{k-1}\| \neq 0$ set

$$\beta_{[u]}^k = (\beta_{[u]}^{k-1} + \gamma_{[u]}^{k-1}) I(\|\beta_{[u]}^{k-1} + \gamma_{[u]}^{k-1}\| \geq \sqrt{h}). \tag{2.25}$$

If $\|\beta_{[u]}^{k-1}\| = 0$ set

$$\beta_{[u]}^k = (\gamma_{[u]}^{k-1}) I(\|\gamma_{[u]}^{k-1}\| > \sqrt{h}). \tag{2.26}$$

As discussed in section 2.5, $\|\beta_{[u]}^{k-1} + \gamma_{[u]}^{k-1}\|$ and $\sqrt{h}$ are real numbers, therefore the event $\|\beta_{[u]}^{k-1} + \gamma_{[u]}^{k-1}\| = \sqrt{h}$ happens with zero probability. Therefore for practical purposes the V-L0LS-CD updates (2.25) and (2.26) are equivalent to,

$$\beta_{[u]}^k = (\beta_{[u]}^{k-1} + \gamma_{[u]}^{k-1}) I(\|\beta_{[u]}^{k-1} + \gamma_{[u]}^{k-1}\| \geq \sqrt{h}). \tag{2.27}$$

Similar to the L0LS-CD algorithm, proper initialization is vital to the V-L0LS-CD algorithm and this will be addressed in section 2.12. V-L0LS-CD can be terminated using similar termination criterion as that of the L0LS-CD algorithm.

2.8 l_0 Penalized Least Squares of Grouped Variables (gL0LS)

As mentioned in section 1.7 some applications require regression of grouped variables. If the coefficient vector is partitioned into m groups $\beta = [\bar{\beta}_1^T, \cdots, \bar{\beta}_j^T, \cdots, \bar{\beta}_m^T]^T$ and the size of the j^{th} partition is p_j, then $\sum_{j=1}^m p_j = p$. Then the linear regression

model (1.1) can be modified to handle grouped variables,

$$y = \sum_{j=1}^{m} X_j \bar{\beta}_j + \varepsilon, \tag{2.28}$$

where X_j is a $n \times p_j$ column wise matrix partition of the regression matrix corresponding to the $\bar{\beta}_j$ coefficient group. Here the X_j partitions are assumed to be orthonormalized, thus $X_j^T X_j = I_{p_j}, j = 1, \cdots, m$.

Sparsity measures discussed in section 1.4 does not promote group sparsity. They can be extended to suite regression of grouped variables as follows,

$$\text{Group} \quad l_0 \quad : \quad |||\beta|||_{r,0} = \sharp\{j,\ \|\bar{\beta}_j\|_r \neq 0\} = \sum I(\|\bar{\beta}_j\|_r \neq 0), \tag{2.29a}$$

$$\text{Group} \quad l_q \quad : \quad |||\beta|||_{r,q} = \left(\sum \|\bar{\beta}_j\|_r^q\right)^{1/q}, 0 < q < 1, \tag{2.29b}$$

$$\text{Group} \quad l_1 \quad : \quad |||\beta|||_{r,1} = \sum \|\bar{\beta}_j\|_r, \tag{2.29c}$$

where $r > 1$ with $r = 2$ being the most common choice. Three algorithms that minimize group l_1 penalized least squares with $r = 2$ is given in [278]. Group l_1 sparsity measure with $r = 2$ is also used in [64, 63] and group l_q sparsity measure with $0 < r < 2$ is used in [143]. l_0 penalized least squares criterion (2.2) can be modified to promote group sparsity (gL0LS) as follows,

$$\bar{J}(\beta) = \|y - X\beta\|^2 + h|||\beta|||_{r,0}. \tag{2.30}$$

Next section develops an algorithm to minimize $\bar{J}(\beta)$ with $r = 2$.

2.9 gL0LS-CD

Similar to the extension of the L0LS-CD algorithm to multivariate regression in section 2.7, this section develops its extension to grouped variables which will be referred to as gL0LS-CD(group l_0 penalized least squares via cyclic descent). For

any pair β, β^o the difference in the value of $\bar{J}(\beta)$ is given by,

$$\begin{aligned}\bar{J}(\beta) =&\bar{J}(\beta^o) - 2(\beta-\beta^o)^T\gamma^o + (\beta-\beta^o)^T X^T X(\beta-\beta^o)\\ &+ h\sum_1^m I(\|\bar{\beta}_j\|\neq 0) - h\sum_1^m I(\|\bar{\beta}_j^o\|\neq 0).\end{aligned} \tag{2.31}$$

where $\gamma^o = X^T(y - X\beta^o)$. Unlike L0LS-CD, gL0LS-CD updates a group of coefficients at each iteration. Thus given the iterate k, factor $k = lm + u$ where l is an integer and $1 \leq u \leq m$. From (2.31) the change in $\bar{J}(\beta)$ at consecutive iterations is given by,

$$\begin{aligned}\Delta\bar{J}_k =&\bar{J}(\beta^k) - \bar{J}(\beta^{k-1})\\ =& -2\bar{\Delta}_u^{k,T}\bar{\gamma}_u^{k-1} + \|\bar{\Delta}_u^k\|^2 + hI(\|\bar{\beta}_u^k\|\neq 0) - hI(\|\bar{\beta}_u^{k-1}\|\neq 0),\end{aligned} \tag{2.32}$$

where $\bar{\gamma}_u^{k-1} = X_u^T(y - X\beta^{k-1})$ and $\bar{\Delta}_u^k = \bar{\beta}_u^k - \bar{\beta}_u^{k-1}$. Repeating the argument given in section 2.5, the gL0LS-CD update becomes,

If $\|\bar{\beta}_u^{k-1}\| \neq 0$ set

$$\bar{\beta}_u^k = (\bar{\beta}_u^{k-1} + \bar{\gamma}_u^{k-1})I(\|\bar{\beta}_u^{k-1} + \bar{\gamma}_u^{k-1}\| \geq \sqrt{h}). \tag{2.33}$$

If $\|\bar{\beta}_u^{k-1}\| = 0$ set

$$\bar{\beta}_u^k = (\bar{\gamma}_u^{k-1})I(\|\bar{\gamma}_u^{k-1}\| > \sqrt{h}). \tag{2.34}$$

As discussed in sections 2.5 and 2.7, $\|\bar{\beta}_u^{k-1} + \bar{\gamma}_u^{k-1}\|$ and $\sqrt{h}$ are real numbers, therefore the event $\|\bar{\beta}_u^{k-1} + \bar{\gamma}_u^{k-1}\| = \sqrt{h}$ happens with zero probability. Therefore for practical purposes the gL0LS-CD updates (2.33) and (2.34) are equivalent to,

$$\bar{\beta}_u^k = (\bar{\beta}_u^{k-1} + \bar{\gamma}_u^{k-1})I(\|\bar{\beta}_u^{k-1} + \bar{\gamma}_u^{k-1}\| \geq \sqrt{h}). \tag{2.35}$$

The gL0LS-CD algorithm is used in the sparse network topology application discussed in chapter 4 where it will be compared with group LASSO.

2.10 Stability Analysis

The convergence analysis is far from straightforward. The criterion is not convex so no easy analysis is possible. More generally the global convergence theorem of [148, 137, 125] fails because $J(\beta)$ is not continuous in β nor is the update (2.13) closed. The methods of [247] also do not apply for related reasons. Also the results of [153] do not apply since they require the penalty be differentiable. So something different is needed and the approach of [252] is followed to some extent.

Although the development of the L0LS-CD algorithm supports both over-determined and under-determined systems, the stability analysis given here applies only to over-determined systems. The simulation results given in section 2.11.5 will show that the L0LS-CD algorithm can be successfully applied to under-determined systems. Thus the stability analysis of the L0LS-CD algorithm on under-determined systems is open for future research.

This analysis proceeds in several stages. First the fixed points of (2.13) are identified. Then a fundamental descent lemma is developed and the convergence analysis then follows.

For the L0LS-CD update (2.13), denote the set of limit points by Γ_L, the set of fixed points by Γ_F and the set of stationary points by Γ_S.

2.10.1 Stationary Points of L0LS

Lemma L1. The stationary points of L0LS are isolated.

Proof. This is a consequence of Theorem I(c) as follows. If two stationary points have the same non-zero set Γ_c then since the solution to $\gamma_u = 0, u \in \Gamma_c$ is unique the two stationary points coincide.

If two stationary points differ in at least one zeroed coefficient then the non-zeroed coefficient obeys (Ic) and so the two stationary points cannot be connected.

There are no other possibilities and the result follows.

2.10.2 Fixed Points of L0LS-CD

The fixed points are obtained by setting $\beta_u^k = \beta_u^{k-1}$ in (2.13), denote fixed points of L0LS-CD as β^f, yielding,

if $\beta_u^f \neq 0$ set

$$\beta_u^f = (\beta_u^f + \gamma_u^f) I(|\beta_u^f + \gamma_u^f| \geq \sqrt{h}), \tag{2.36}$$

if $\beta_u^f = 0$ set

$$\beta_u^f = \gamma_u^f I(|\gamma_u^f| > \sqrt{h}) = 0. \tag{2.37}$$

where $\gamma_u^f = x_{(u)}^T (y - X\beta^f)$. Then for $1 \leq u \leq p$,

(a) From (2.37), $u \in \Gamma_0$ iff, $|\gamma_u^f| \leq \sqrt{h}$.

(b) From (2.36), $u \in \Gamma_c$ iff, $\beta_u^f = (\beta_u^f + \gamma_u^f) \Rightarrow \gamma_u^f = 0$,

(c) From (2.36), $u \in \Gamma_c$ iff, $|\beta_u^f + \gamma_u^f| \geq \sqrt{h} \Rightarrow |\beta_u^f| \geq \sqrt{h}$.

Thus any finite limit point of L0LS-CD obeys the optimality conditions of Theorem I.

We have thus established:

Lemma L2. The fixed points are stationary points i.e. $\Gamma_F \subseteq \Gamma_S$.

2.10.3 Descent Lemma

Write (2.14) and (2.15) compactly as,

$$\Delta J_k = -D(\beta_u^{k-1}, \gamma_u^{k-1}), \tag{2.38}$$

where $D(\beta_u^0, \gamma_u^0) \geq 0$ is given by,

(Da) if $\beta_u^0 \neq 0$

$$D(\beta_u^0, \gamma_u^0) = (\gamma_u^0)^2 + [h - (\beta_u^0 + \gamma_u^0)^2] I(|\beta_u^0 + \gamma_u^0| < \sqrt{h}),$$

(Db) if $\beta_u^0 = 0$

$$D(\beta_u^0, \gamma_u^0) = [(\gamma_u^0)^2 - h]I(|\gamma_u^0| > \sqrt{h}).$$

We now have the following fundamental property.

Lemma D. $D(\beta_u^0, \gamma_u^0) \geq 0$ and if $\beta_u^1 = A(\beta_u^0, \gamma_u^0)$ then $D(\beta_u^0, \gamma_u^0) = 0 \Rightarrow \beta_u^1 = \beta_u^0$.

proof The fact that $D \geq 0$ follows by inspection of the definition. For the main part of the lemma there are two cases.

Case I : $\beta_u^0 \neq 0$.
From Da we obtain $\gamma_u^0 = 0$ and either :
(Ia) $I_{u,<} = 0 \equiv I_{u,\geq} = 1$ or
(Ib) $I_{u,<} = 1$ and $(\beta_u^0 + \gamma_u^0)^2 = h$.
For (Ia), $|\beta_u^0 + \gamma_u^0| \geq \sqrt{h}$ so $|\beta_u^0| \geq \sqrt{h}$. Also we get from the update $\beta_u^1 = \beta_u^0 + \gamma_u^0 = \beta_u^0$.
For (Ib), $|\beta_u^0 + \gamma_u^0| < \sqrt{h}$ which contradicts $(\beta_u^0 + \gamma_u^0)^2 = h$.
So only (Ia) can occur and we get $\beta_u^1 = \beta_u^0$ as required.

Case II : $\beta_u^0 = 0$.
From Db we obtain either:
(IIa) $|\gamma_u^0| \leq \sqrt{h}$ or
(IIb) $|\gamma_u^0| > \sqrt{h}$.
For (IIa) the update gives $\beta_u^1 = 0 = \beta_u^0$.
For (IIb) we also get $|\gamma_u^0| = h$ which is a contradiction so that (IIb) cannot occur.
So we can only have (IIa) and so we get $\beta_u^1 = \beta_u^0$ as required.

2.10.4 Boundedness

We introduce a rank condition.

Condition R. $X^T X$ has full rank.

Lemma L3. $\Delta J^k = D_k \to 0$ and under condition R, $\|\beta^k\|$ is bounded.

Proof. Iterating the descent lemma gives $0 \leq J^k \leq J(\beta^0) < \infty$. Thus J^k is a bounded non-increasing sequence and so must have a finite limit say J^∞. Thus

$D_k = \Delta J^k = J^k - J^{k-1} \to J^\infty - J^\infty = 0$. Also J^k bounded $\Rightarrow$ β^k bounded provided condition R holds.

2.10.5 Convergence of Iterate Differences

Lemma L4. Under condition R, $\beta_u^k - \beta_u^{k-1} \to 0$.

Proof. We apply Lemma L3. Set $S_u^{k-1} = \beta_u^{k-1} + \gamma_u^{k-1}$. We have $D_k \to 0$ and so

$$I(\beta_u^{k-1} \neq 0)[(\gamma_u^{k-1})^2 + [h - (S_u^{k-1})^2] I(|S_u^{k-1}| < \sqrt{h}) \to 0$$
$$I(\beta_u^{k-1} = 0)[(\gamma_u^{k-1})^2 - h] I(|\gamma_u^{k-1}| > \sqrt{h}) \to 0$$

There are two cases.

Case I: $I(\beta_u^{k-1} = 0) \to 0$ and case II: $I(\beta_u^{k-1} = 0) \to 1$. We consider each in turn.

Case I. $I(\beta_u^{k-1} = 0) \to 0$.

Now given $\epsilon > 0$ we can find k_o so that for all $k \geq k_o, I(\beta_u^{k-1} = 0) < \epsilon$. But this means $I(\beta_u^{k-1} = 0) = 0$ for all $k \geq k_o$. So $\beta_u^{k-1} \neq 0$ for all $k \geq k_o$. So $I(\beta_u^{k-1} \neq 0) = 1$ for all $k \geq k_o$. Thus $(\gamma_u^{k-1})^2 \to 0$ and either $(S_u^{k-1})^2 \to h$ or $I(|S_u^{k-1}| < \sqrt{h}) \to 0$. But since $(\gamma_u^{k-1})^2 \to 0$ this means either $(\beta_u^{k-1})^2 \to h$ or $I(|\beta_u^{k-1}| < \sqrt{h}) \to 0$.

In the latter sub-case we can find $k_1 \geq k_o$ so that for all $k \geq k_1, I(|\beta_u^{k-1}| < \sqrt{h}) = 0$. Then for all $k \geq k_1, \beta_u^k = \beta_u^{k-1} + \gamma_u^{k-1}$ and so $\beta_u^k - \beta_u^{k-1} \to 0$.

In the other sub-case there is a problem if $|\beta_u^{k-1}|$ increases to $\sqrt{h}$. But we now show this cannot occur. If e.g. $|\beta_u^{k_a - 1}| < \sqrt{h}$ then $|\beta_u^{k_a}| = 0$ since $\gamma_u^{k-1} \to 0$; but then $\beta_u^{k_a+1} = 0$ and indeed $\beta_u^{k_a+r} = 0$ for all $r \geq 0$. This contradicts $|\beta_u^{k-1}| \to \sqrt{h}$.

We can thus conclude that for some $k_2 \geq k_1$ we have for all $k \geq k_2, |\beta_u^{k-1}| \geq \sqrt{h}$. And so $\beta_u^k = \beta_u^{k-1} + \gamma_u^{k-1}$ and so again $\beta_u^k - \beta_u^{k-1} \to 0$.

Thus the result is established for case I.

Case II. $I(\beta_u^{k-1} = 0) \to 1$.

We can now find k_o so that for all $k \geq k_o, \beta_u^{k-1} = 0$. But then $\beta_u^k - \beta_u^{k-1} = 0$ for all $k \geq k_o$ and the result is established in case II.

The proof is complete.

2.10.6 Limit Points of L0LS-CD

Lemma L5. Under condition R, the limit points of L0LS-CD are fixed points i.e. $\Gamma_L \subseteq \Gamma_F$.

Proof. From Lemma L3 $(\beta_u^{k-1}, \beta_u^k)$ is bounded and so has at least one limit point. Let (β_u^-, β_u^+) be one such limit; then we can find a subsequence k' with $(\beta_u^{k'-1}, \beta_u^{k'}) \to (\beta_u^-, \beta_u^+)$. However by Lemma L4, $\beta_u^{k'} - \beta_u^{k'-1} \to 0$ and so $\beta_u^+ = \beta_u^-$ and so the limit point is a fixed point and the result is established.

2.10.7 Connectedness

Lemma L6. Under condition R, the set of limit points of L0LS-CD is a compact connected set.

Proof. We use Ostrowski's theorem [176]. Namely if $\|\beta^k - \beta^{k-1}\| \to 0$ and β^k is bounded then Γ_L is a connected set. The first part follows from L4 and the second part from L3.

2.10.8 Convergence of Iterates

Theorem II. Under condition R, the iterates converge to a stationary point.

Proof. By L5, $\Gamma_L \subseteq \Gamma_F$. By L2, $\Gamma_F \subseteq \Gamma_S$. By L1, Γ_S consists of isolated points. Thus Γ_L consists of isolated points. But by L6, Γ_L is compact and connected. So Γ_L must consist of a single point. And so the iterates must converge to it.

2.11 L0LS-CD Simulation

Several issues need to be addressed before comparing the performance of L0LS-CD with other existing algorithms. Thus section 2.11.1 discusses performance

measures, section 2.11.2 discuss the choice of h and section 2.11.3 discuss L0LS-CD initialization. Finally section 2.11.4 present simulation setup followed by section 2.11.5 which present simulation results comparing the performance of L0LS-CD with other algorithms.

2.11.1 Performance Measures

A method of evaluating the performance or a measure of the desirability of an estimate has to be established before comparing the performance of sparse regression algorithms. Although sparse regression has been a topic of great interest, very little attention has been paid on performance measures of sparse estimates, especially when the measurement vector is contaminated by noise. Thus there are no universally accepted performance measures.

Establishing performance measures for sparse estimates in noisy systems is not straightforward. In noiseless systems it is reasonable to expect the estimate to be identical to the original coefficient vector. This expectation is practical only on very mildly noisy systems because information gets drowned in noise.

Goals of sparse approximation are given in section 1.3. Expecting accuracy and sparsity at the same time is tricky because a little sacrifice in one aspect normally tends to improve the other. Thus the choice of the performance measure will depend on the requirements of the application. Applications based on prediction will be more concerned about the estimated signal $\hat{\mu} = X\hat{\beta}$ rather than the atoms in the active set of $\hat{\beta}$. In contrast applications such as topology identification of a sparse network or identifying active sensors of a sensor network will be more concerned about the model of the estimate.

Thus this thesis uses three kinds of performance measures to compare the performance of sparse regression algorithms. Given h let us denote the estimate as $\hat{\beta}_h$ and the true coefficient vector as $\beta^\star$.

(i) Signal mean squared error (MSE_μ),

$$\text{MSE}_\mu = \|X(\hat{\beta}_h - \beta^\star)\|^2 / \|X\beta^\star\|, \tag{2.39}$$

This is relevant to prediction.

(ii) Parameter mean squared error (MSE_β),

$$\text{MSE}_\beta = \|\hat{\beta}_h - \beta^\star\|^2 / \|\beta^\star\|^2. \tag{2.40}$$

This is relevant to estimating β.

(iii) Selection of the correct model.

Introduce $\Gamma_0 = \{j : \beta_j^\star = 0\}$, $\Gamma_c = \{j : \beta_j^\star \neq 0\}$, $\hat{\Gamma}_0 = \{j : \hat{\beta}_{h,j} = 0\}$ and $\hat{\Gamma}_c = \{j : \hat{\beta}_{h,j} \neq 0\}$. Thus we can define,

$$\begin{aligned}
&\text{Number of true positives (TP)} && = \dim(\Gamma_c \bigcap \hat{\Gamma}_c) \\
&\text{Number of false positives (FP)} && = \dim(\Gamma_0 \bigcap \hat{\Gamma}_c) \\
&\text{Number of false negatives (FN)} && = \dim(\Gamma_c \bigcap \hat{\Gamma}_0) \\
&\text{Number of true negatives (TN)} && = \dim(\Gamma_0 \bigcap \hat{\Gamma}_0)
\end{aligned}$$

where dim( ) represent the dimension of the set. Now we can define,

$$\text{True positive rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.41}$$

$$\text{False positive rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.42}$$

which are important performance measures depicting the accuracy of the selected model.

Sparsity of the estimate and the speed of the algorithm are also major considerations when analyzing the effectiveness of an algorithm.

2.11.2 Selection of h

While the performance measures presented in section 2.11.1 can be calculated in a simulation, they cannot be computed in practice because $\beta^\star$ is unknown. Had this been a possibility we could have selected h such that the estimate gives the optimum value of a selected performance measure.

Proper selection of h is widely neglected in the literature. [20] have dismissed pIHT as inferior compared to cIHT as a result of improper selection of the penalty

parameter, as demonstrated below in section 2.11.5. [116] sets $h = 3$ for a range of signal to noise ratios (SNR) and [187, 40] used the L-curve method to select the penalty parameter. Since the L-curve method has been severely criticized in [96, 260], it is not considered in this thesis.

The value of h determines the sparsity of the estimate and therefore its model. An introduction to model selection criteria was given in section 1.8. Although SURE based criteria have been developed to select h in an l_1 penalized least squares problem, no such method exists for algorithms that use other sparsity measures.

Bayesian information criterion (BIC) [200, 135, 123] is a model selection criterion that is used widely but it is generally used to select discrete tuning parameters. Since the value of h determines the number of non-zero coefficients retained by the algorithm, this enables us to use BIC to select h. When noise variance σ^2 is known,

$$\min_h \text{BIC} = \frac{\|y - X\hat{\beta}_h\|^2}{\sigma^2} + r\ln(n), \tag{2.43}$$

where r is the number of non-zero coefficients of $\hat{\beta}_h$ i.e. $r = \dim(\hat{\Gamma}_c)$. If the variance of noise is unknown then,

$$\min_h \text{BIC} = \ln\left(\frac{\|y - X\hat{\beta}_h\|^2}{n}\right) + \frac{r}{n}\ln(n), \tag{2.44}$$

Denote the minimizer of BIC by $\hat{h}$. It is then more logical to compare the performance using,

$$\begin{aligned}
\text{MSE}_\mu &= E[\|X(\hat{\beta}_{\hat{h}} - \beta^\star)\|^2/\|X\beta^\star\|^2], \\
\text{MSE}_\beta &= E[\|\hat{\beta}_{\hat{h}} - \beta^\star\|^2/\|\beta^\star\|^2].
\end{aligned}$$

These MSE's are estimated by averaging over many repeats. Similarly we can redefine the sets needed to calculate the TPR and FPR as, $\hat{\Gamma}_0 = \{j : \hat{\beta}_{\hat{h},j} = 0\}$ and $\hat{\Gamma}_c = \{j : \hat{\beta}_{\hat{h},j} \neq 0\}$.

It should be noted that BIC is not suitable for model selection when the model

space is large. Instances where BIC fails are shown in chapter 6 and in [251]. The BIC criterion has been extended to handle large models in [33]. However in the simulations presented in this thesis BIC worked well, thus method developed in [33] was not employed.

2.11.3 L0LS-CD Initialization

Since L0LS-CD terminates at a local minimum of $J(\beta)$, the initialization of the algorithms has a significant impact on its performance. The following four types of initializations were considered and their impact on over-determined and under-determined systems were analyzed.

(i) All zero initialization ($\beta^0 = 0$).

(ii) Random starting vectors [i.e. Gaussian with zero mean and unit variance].

(iii) OMP estimate.

(iv) L1LS (l_1 penalized least squares) estimate.

L0LS-CD estimates for over-determined systems seem to depend very little on the initialization. This is especially true for high sparsity levels. Extensive simulations suggest that all zero initialization is the best choice for over-determined systems.

Unlike over-determined systems the estimates of under-determined systems seem to depend heavily on the choice of initialization. Extensive simulations suggest that initialization with L1LS estimates produces the best results.

2.11.4 Scalar Regression Simulation Setup

For the scalar regression simulations the data was generated as follows. The entries of the X matrix are independent Gaussians with zero mean and unit variance. The columns of X were scaled to have unit norm $\|x_{(j)}\| = 1, j = 1, \ldots, p$.

Denote the number of nonzero coefficients as r, then sparsity $= 1 - r/p$, for over-determined systems and sparsity $= 1 - r/n$, for under-determined systems.

It is important to note that sparsity is lower bounded by 0.5 for under-determined noiseless systems [90]. Since the estimate of a noisy system cannot perform better than that of a noiseless system, range of sparsity was limited to 0.5 - 1 for under-determined systems. The coefficient vector $\beta^\star$ was generated by placing r non-zero values at random locations in a p dimensional vector. The non-zero values were selected using one of three different probability distributions as follows,

(i) Gaussian distribution with zero mean and unit variance.

(ii) Laplace distribution with the zero mean and diversity set to 2.

(iii) Bernoulli distribution with outcomes [1,-1] with probability of each event set to 0.5.

The y vector was calculated from (1.1), where the noise vector ε is Gaussian with zero mean and variance σ^2. The variance of the noise depends on the SNR,

$$\text{SNR} = \frac{\|X\beta^\star\|^2}{n\sigma^2}. \tag{2.45}$$

Note that here we do not use the traditional definition of SNR as the logarithmic decibel scale is not used. For a given X and $\beta^\star$ an initial set of simulations were done to select h using BIC (2.44) for each algorithm. Then a separate set of simulations were done to compare the performance of the algorithms. At this stage the selected value of h was kept fixed at each iteration unlike in [187].

L0LS-CD was compared with L1LS (estimate of which is the same as that of LASSO since l_1 is convex), regularized FOCUSS [187], pIHT [20], cIHT (over-determined [20], under-determined [21]) OMP and CoSaMP [166]. L1LS was optimized using cyclic descent as in [74, 278]. As recommended by the authors of [187] q was set to 0.8 for regularized FOCUSS.

The development of the IALZ algorithm presented in [115, 117] does not support noise. However systems with noise are considered in the simulations of [115, 117] and the noise is handled in an ad-hoc manner by adjusting the algorithm termination threshold according to the SNR, thus IALZ is not included in the preceding simulations. Although the vector version, JLZA [116] supports noise, when the vector length was set to 1 its performance degraded dramatically.

Its performance may be improved with a different configuration of its tuning parameters but this is a separate study by itself and will not be attempted in this thesis.

As mentioned in chapter 1, CoSaMP [166] cannot support low sparsity levels. If r is the expected number of non-zero entries in the estimate, CoSaMP requires finding the pseudoinverse of a sub matrix of X which can have up to $3r$ columns. In order for this sub matrix of X to have full rank in under-determined systems, r has to be less than $n/3$. Denote the residual as e^k, CoSaMP calculates a signal proxy (vector of residual correlations), $X^T e^k$ and then selects the indices of the $2r$ largest-magnitude components. For this to be possible in over-determined systems r has to be less than $p/2$.

As recommended in [20], pIHT was initialized with the output of OMP, $\beta^0 = \beta_{OMP}$. cIHT was initialized as $\beta^0 = \beta_{OMP}$ for over-determined systems and as $\beta^0 = 0$ for under-determined systems as recommended by [20] and [21] respectively. L0LS-CD was initialized with all zeros ($\beta^0 = 0$) and with the output of L1LS ($\beta^0 = \beta_{l_1}$) to show the importance of proper initialization.

2.11.5 L0LS-CD Performance Comparison

The algorithms were compared on over-determined as well as under-determined systems. For each system, n, p and SNR were first kept fixed and r was varied to show the performance with varying sparsity and then n, p and r were kept fixed and SNR was varied to show the performance with varying noise levels. X is kept fixed throughout the simulation. For each sparsity level 20 β vectors were generated and for each β vector 50 y vectors were generated. The median of the results were considered for performance comparison.

Under-determined Systems: First consider under-determined systems. Set $n = 50$, $p = 128$, SNR $= 10$ and vary r from 5 to 25 (CoSaMP cannot support $r > 16$). Refer figure 2.1 for the performance of the algorithms with varying sparsity when the non-zero values of $\beta^\star$ were drawn from a Gaussian distribution. Regularized FOCUSS and L1LS has the highest TPR. However they also have a very high FPR. So these algorithms produce estimates with low sparsity and they are unable to identify the correct model, thus undesirable. L0LS-CD

with $\beta^0 = \beta_{l_1}$ produce the next highest TPR while maintaining the lowest FPR. Regularized FOCUSS and L1LS produces the lowest MSE_μ and MSE_β however as discussed above their estimates have very low sparsity. L0LS-CD with $\beta^0 = \beta_{l_1}$ has the next lowest MSE_μ and MSE_β particularly at low sparsity levels. Its very clear from all the plots that L0LS-CD with $\beta^0 = \beta_{l_1}$ is far superior to L0LS-CD with $\beta^0 = 0$. Thus it is clear that proper initialization is very important.

Refer figures 2.2 and 2.3 for the performance of the algorithms with varying sparsity when the non-zero values of $\beta^\star$ were drawn from Laplace and Bernoulli distributions respectively. The performance of the algorithms in figure 2.2 seem identical to that of figure 2.1. As shown in figure 2.3 the performance of OMP, pIHT and CoSaMP seem to degrade when the non-zero values of $\beta^\star$ were drawn from a Bernoulli distribution. However L0LS-CD with $\beta^0 = \beta_{l_1}$ continues to outperform the others.

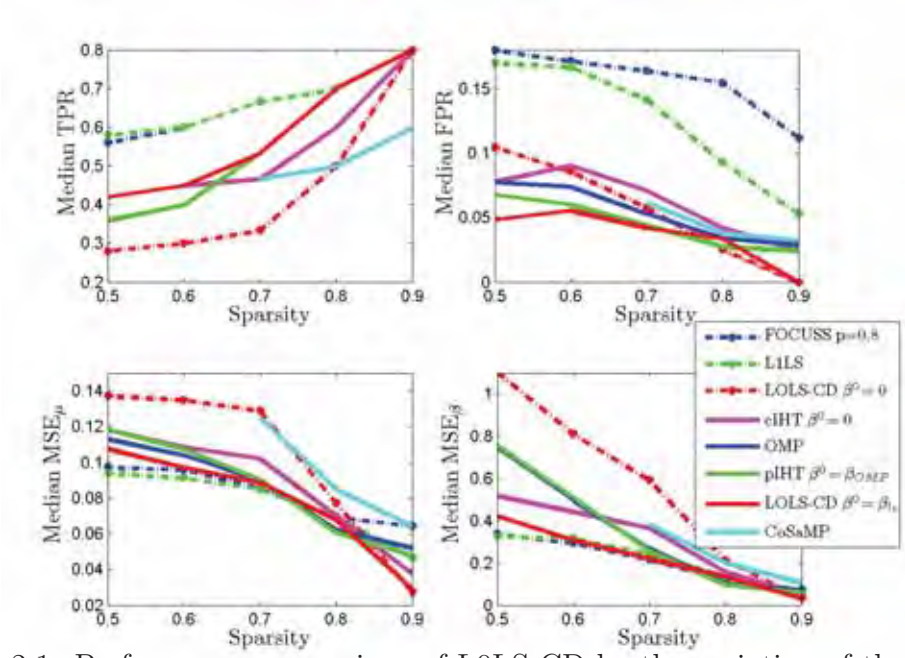

Figure 2.1: Performance comparison of L0LS-CD by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128, SNR = 10 and non-zero values of $\beta^\star$ drawn from a Gaussian distribution.

To compare performance of algorithms at various noise levels, set $n = 50$, $p = 128$, $r = 15$ and vary SNR from 30 to 3. Refer figure 2.4 for the results when the non-zero values of $\beta^\star$ were drawn from a Gaussian distribution. Similar to

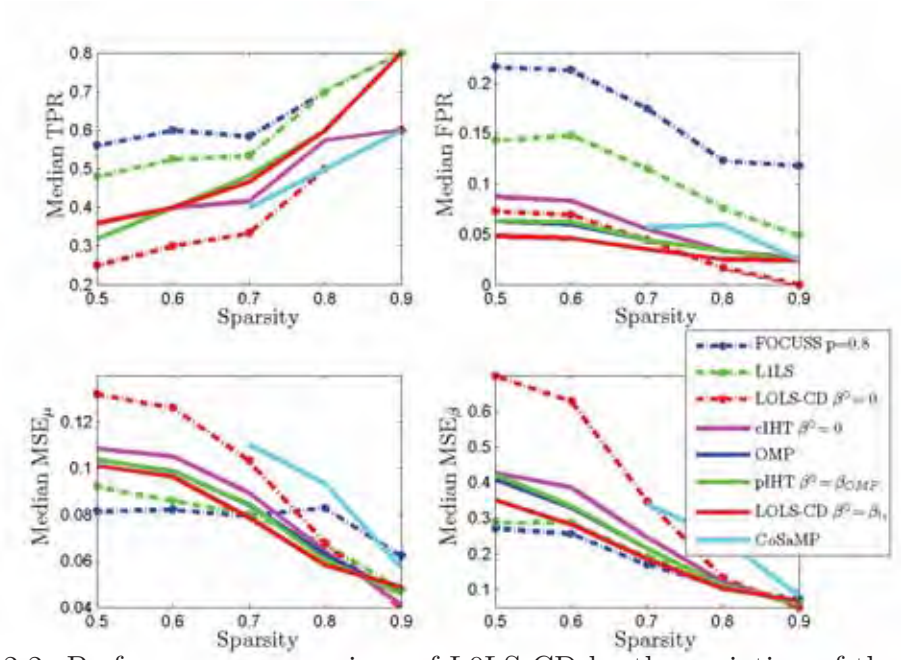

Figure 2.2: Performance comparison of L0LS-CD by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128, SNR = 10 and non-zero values of $\beta^\star$ drawn from a Laplace distribution.

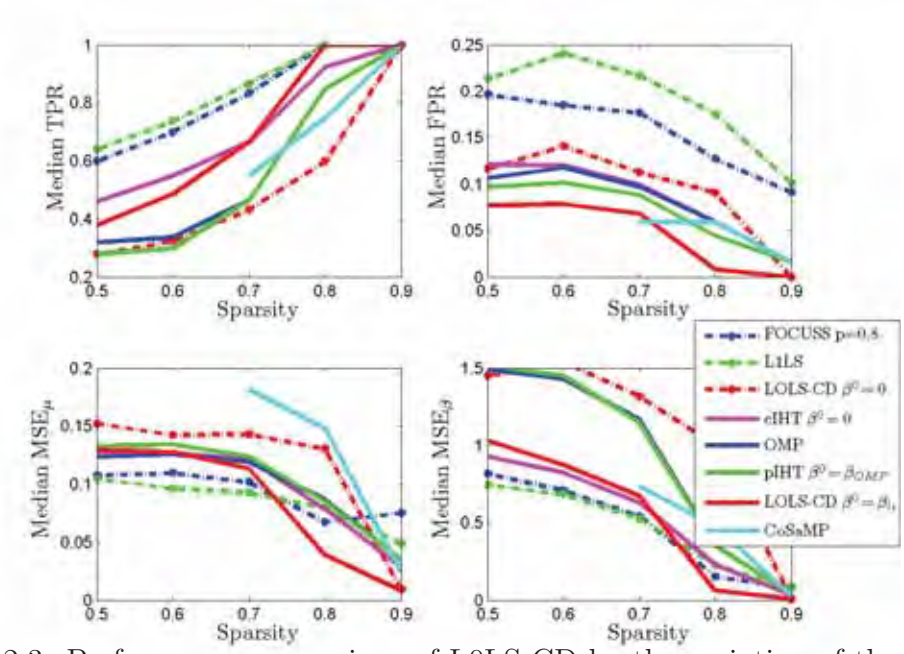

Figure 2.3: Performance comparison of L0LS-CD by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128, SNR = 10 and non-zero values of $\beta^\star$ drawn from a Bernoulli distribution.

figure 2.1, although regularized FOCUSS and L1LS has the highest TPR it also has the highest FPR and thus they are undesirable. L0LS-CD with $\beta^0 = \beta_{l_1}$ has the second highest TPR while maintaining the lowest FPR, MSE_μ and MSE_β.

When an Laplace distribution was used to generate the $\beta^\star$ vector, the results seem identical to figure 2.4. Similar to the observation in figure 2.3, the performance of OMP, pIHT and CoSaMP degraded when a Bernoulli distribution was used to generate the $\beta^\star$ vector. In both cases L0LS-CD with $\beta^0 = \beta_{l_1}$ outperformed the others. Refer appendix 2.B for the figures showing the performance of the algorithms with varying SNR when the $\beta^\star$ vector was generated using Laplace and Bernoulli distributions.

Although the authors of [20] state that cIHT has superior exact recovery capabilities compared to pIHT, it is apparent from figures 2.1, 2.2 and 2.4 that when h is selected properly pIHT has lower FPR than that of cIHT while maintaining similar TPR. Thus when h is selected properly pIHT has the potential of producing superior results to cIHT in terms of exact recovery. However when Bernoulli distribution is used to generate the $\beta^\star$ vector the performance of pIHT drops dramatically as shown in figure 2.3.

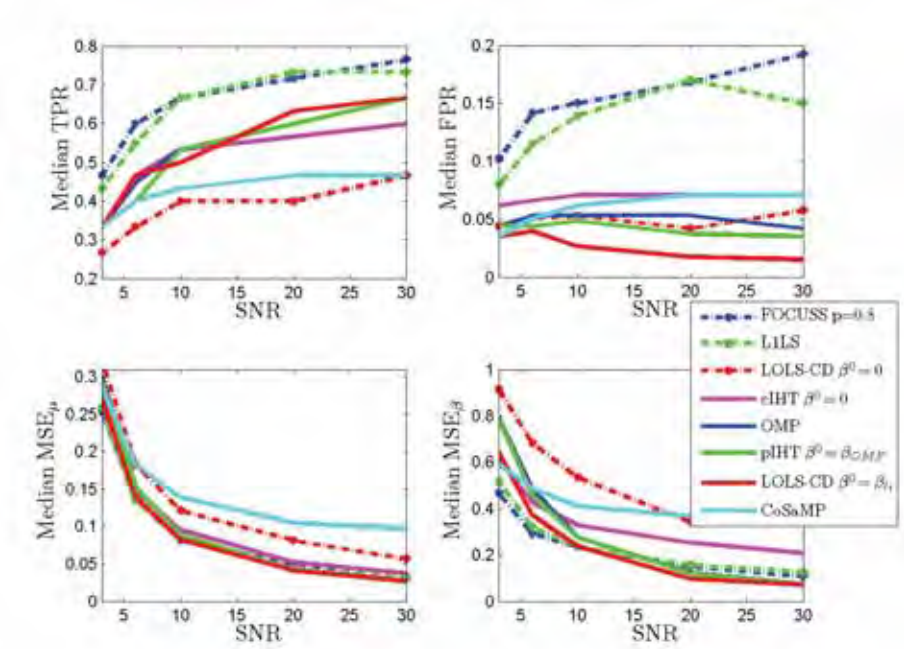

Figure 2.4: Performance comparison of L0LS-CD by the variation of the performance measures as a function of SNR in an under-determined system with n = 50, p = 128, r = 15 and non-zero values of $\beta^\star$ drawn from a Gaussian distribution.

Over-determined Systems: To observe the performance of the algorithms at varying levels of sparsity, set $n = 128$, $p = 50$, SNR $= 10$ and vary r from 5 to 40 (CoSaMP cannot support $r > 25$). To observe the performance of the algorithms at various levels of noise, set $n = 128$, $p = 50$, $r = 15$ and vary SNR from 30 to 3. Figures 2.5 and 2.6 show the performance of the algorithms at various levels of sparsity and SNR when the $\beta^\star$ vector was generated using a Gaussian distribution. It is clear from figures 2.5 and 2.6 that the properties of the algorithms when applied to over-determined systems is noticeably different from that of under-determined systems. Similar to under-determined systems JLZA, regularized FOCUSS and L1LS all produce undesirable, low sparse estimates with very high FPR. CoSaMP produce estimates with high MSE_μ and MSE_β. However unlike in under-determined systems L0LS-CD, pIHT, cIHT and OMP produce similar results. These four algorithms seem to be compatible when applied to over-determined systems, but their performance is clearly different when applied to under-determined systems. L0LS-CD, pIHT, cIHT and OMP produce estimates with lowest MSE_μ and MSE_β while maintaining the lowest FPR and moderately high TPR. This again emphasizes the fact that pIHT and cIHT can produce comparable results when h is selected properly. All the algorithms performed in a similar manner when Laplace and Bernoulli distributions were used to generate the $\beta^\star$ vector and these results are given in appendix 2.B.

2.12 V-L0LS-CD Simulation

The performance measures of scalar sparse regression discussed in section 2.11.1 can be easily extended to multivariate regression. Given the value of h denote the estimate as $\hat{B}_h$ and the original coefficient matrix as $B^\star$, then

$$\text{MSE}_\mu = E\left(\frac{\sum_{j=1}^{d}\|X(\hat{\beta}_{(j),\hat{h}} - \beta^\star_{(j)})\|^2}{\sum_{j=1}^{d}\|X(\beta^\star_{(j)})\|^2}\right),$$

$$\text{MSE}_\beta = E\left(\frac{\sum_{j=1}^{d}\|\hat{\beta}_{(j),\hat{h}} - \beta^\star_{(j)}\|^2}{\sum_{j=1}^{d}\|\beta^\star_{(j)}\|^2}\right),$$

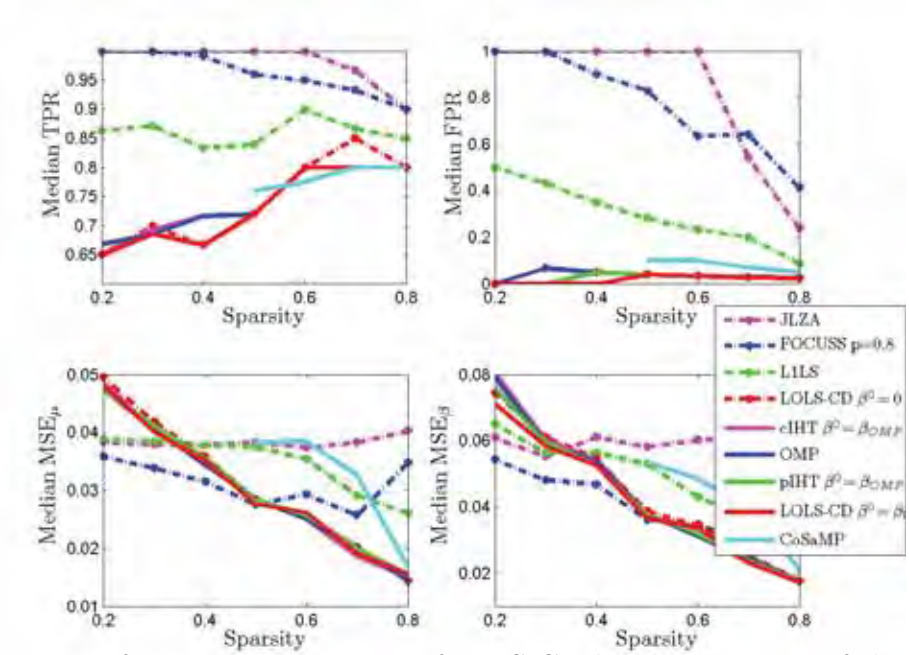

Figure 2.5: Performance comparison of L0LS-CD by the variation of the performance measures as a function of sparsity in an over-determined system with n = 128, p = 50, SNR = 10 and non-zero values of $\beta^\star$ drawn from a Gaussian distribution.

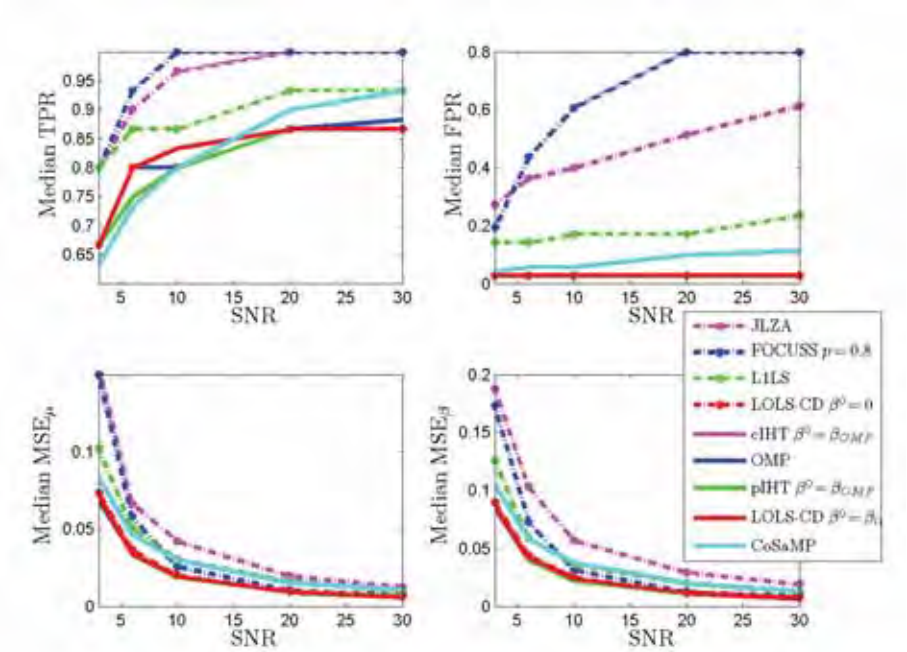

Figure 2.6: Performance comparison of L0LS-CD by the variation of the performance measures as a function of SNR in an over-determined system with n = 128, p = 50, r = 15 and non-zero values of $\beta^\star$ drawn from a Gaussian distribution.

where $\hat{\beta}_{(j),h}$ and $\beta^{\star}_{(j)}$ are the j$^{\text{th}}$ columns of $\hat{B}_h$ and $B^{\star}$ respectively. Similarly the sets needed to calculate TPR and FPR can be redefine as, $\Gamma_0 = \{j : \|\beta^{\star}_{[j]}\| = 0\}$, $\Gamma_c = \{j : \|\beta^{\star}_{[j]}\| \neq 0\}$, $\hat{\Gamma}_0 = \{j : \|\hat{\beta}_{[j],\hat{h}}\| = 0\}$ and $\hat{\Gamma}_c = \{j : \|\hat{\beta}_{[j],\hat{h}}\| \neq 0\}$. Similar to section 2.11.2 $\hat{h}$ is selected using a modification of the BIC criterion as follows,

$$\hat{h} = \min_{h} \text{BIC} = \frac{\sum_{j=1}^{d} \|y(j) - X\hat{\beta}_{(j),h}\|^2}{\sigma^2} + \dot{r}d\ln(nd), \tag{2.46}$$

where $\dot{r}$ is the number of non-zero rows in $\hat{B}$, i.e. $\dot{r} = \dim(\hat{\Gamma}_c)$. Here the noise vectors (columns of matrix E) are considered to be independent of each other (correlation matrix $=\sigma^2 I$).

After extensive simulations, similar to that discussed in section 2.11.3, the best initialization for V-L0lS-CD was found to be the minimizer of V-L1LS criterion.

2.12.1 Multivariate Regression Simulation Setup

For all the multivariate regression simulations the data was generated similar to that of scalar regression. Here only under-determined systems are considered since as shown in section 2.11.5 estimation in an over-determined system is more straightforward.

The dictionary X is created as given in section 2.11.4. Sparsity of B for an under-determined system is $1 - \dot{r}/n$, where $\dot{r}$ is the number of non-zero rows of B. The locations of the non-zero rows were selected from a discrete uniform distribution and the non-zero rows were created by entries from a Gaussian random variable with zero mean and unit variance. In scalar regression simulations, we investigated how the performance of algorithms get influenced by the probability distribution of the non-zero entries of the coefficient vector. As shown in the simulation results of section 2.11.5 this did not have a considerable effect on algorithm performance. Thus only Gaussian distribution is considered in the multivariate regression simulations.

For a given X, B and SNR value the Y was generated from (2.20), where E contain noise vectors of zero mean and variance σ^2. σ^2 depends on the SNR level and the noise vectors are assumed to be independent from each other (correlation

matrix $=\sigma^2 I$).

$$\text{SNR} = \frac{\sum_{j=1}^{d} \|X\beta^{\star}_{(j)}\|^2}{nd\sigma^2}.$$

Similar to the scalar simulation setup 2.11.4 a preliminary set of simulations were done to get $\hat{h}$ by BIC. $\hat{h}$ is then used in a second set of simulations to study the algorithm performance which is kept fixed at each iteration within an algorithm unlike in [40].

V-L0LS-CD was compared with vector l_1 penalized least squares (V-L1LS) [144], regularized M-FOCUSS [40], JLZA [116](tuning parameter settings recommended in [116] were used) and SOMP [245]. V-L1LS is introduced in [144], but is solved by second order cone programming; instead cyclic descent (V-L1LS-CD) was used here. Since the criterion is convex, both algorithms will produce the same answer. As stated in [40], p was set to 0.8 and at the end of the algorithm, Y was orthogonally projected on to the atoms selected by the algorithm. [40] perform hard thresholding of the estimates of regularized M-FOCUSS so that the sparsity of the estimates would equal that of the original B matrix. This step was omitted here as the sparsity of the original B matrix is generally unknown. Similar to the simulations performed in section 2.11, V-L0LS-CD was initialized with all zeros $B^0 = 0$ as well as with the estimate of V-L1LS-CD $B^0 = B_{l_1}$ to show the importance of initialization.

2.12.2 V-L0LS-CD Performance Comparison

For the simulations discussed in this section dimensions similar to [40]; $n = 20$, $p = 30$ were used. X was kept fixed throughout the simulation.

First the variation of the performance measures with sparsity were investigated. Set $d = 3$ and SNR= 10 and vary k from 2 to 10. For each sparsity level 50 B matrices were generated and using each B matrix 100 Y matrices were generated. Results are given in figure 2.7.

Similar to L1LS in scalar regression simulations, V-L1LS-CD has the highest TPR and the highest FPR. This means that V-L1LS-CD like L1LS produces estimates with very low sparsity and is thus undesirable. V-L0LS-CD with $B^0 =$

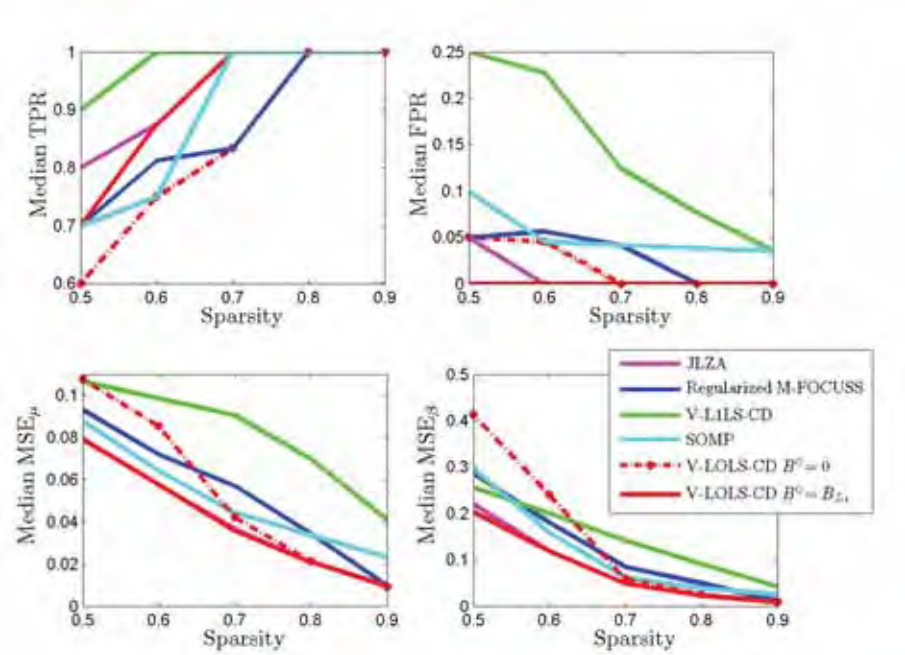

Figure 2.7: Performance comparison of V-L0LS-CD by the variation of the performance measures as a function of sparsity in a multivariate under-determined system with n = 20, p = 30, d = 3 and SNR = 10.

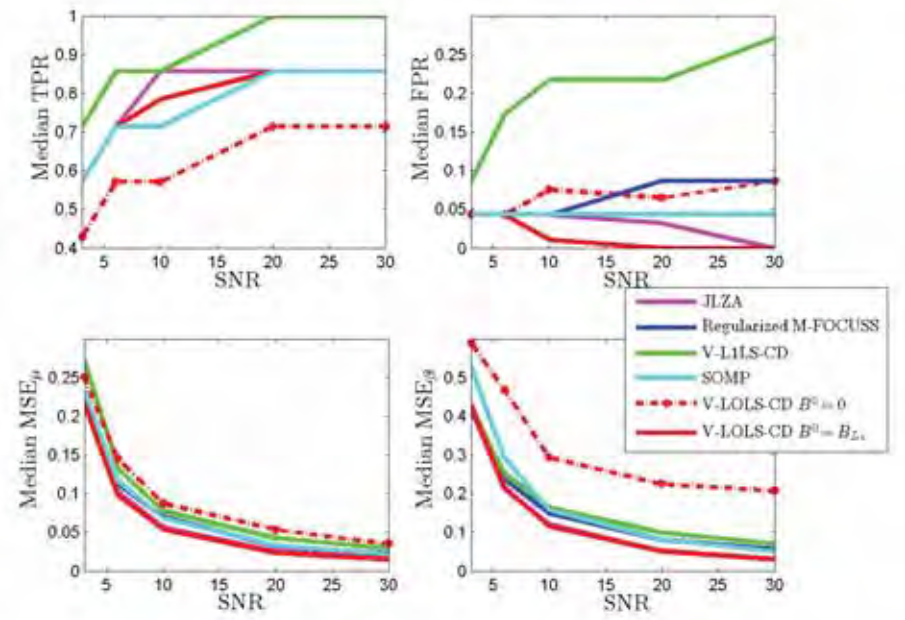

Figure 2.8: Performance comparison of V-L0LS-CD by the variation of the performance measures as a function of SNR in a multivariate under-determined system with n = 20, p = 30, d = 2 and $\dot{r} = 7$.

B_{l_1} has the next highest TPR while maintaining the lowest FPR. Furthermore V-L0LS-CD with $B^0 = B_{l_1}$ has the lowest MSE_μ and MSE_β specially towards the lower sparsity levels. Although JLZA and regularized FOCUSS had poor performance in the scalar regression setting its performance is greatly improved in multivariate regression. However in this example V-L0LS-CD with $B^0 = B_{l_1}$ is superior to the others.

Secondly the variation of the performance measures with SNR were investigated. Set $d = 2$, $k = 7$ and vary SNR from 30 to 3. Results are given in figure 2.8.

Similar to the earlier example V-L1LS-CD has very high FPR and V-L0LS-CD with $B^0 = B_{l_1}$ has the lowest FPR. Furthermore V-L0LS-CD with $B^0 = B_{l_1}$ has the lowest MSE_μ and MSE_β.

From both these examples it is clear that when considered individually V-L1LS-CD produces very low sparsity results with very high FPR and V-L0LS-CD with $B^0 = 0$ produces results with very low TPR. However when V-L0LS-CD is initialized with V-L1LS-CD estimate it produces the best results.

2.13 Conclusion

This chapter discussed exact l_0 denoising and presented a cyclic descent based algorithm (L0LS-CD) to minimize the l_0 penalized least squares criterion. Issues of computational speed were addressed and a greedy method was proposed to enhance speed. Importance of proper initialization, convergence and the stability of the algorithm was also investigated. The impact of the penalty parameter on the performance of the algorithm was illustrated. Simulation results show that L0LS-CD produce superior results in terms of sparsity, MSE_μ, MSE_β and model selection when applied to under-determined systems. In over-determined systems L0LS-CD, pIHT, cIHT and OMP produce comparable results while outperforming JLZA, regularized FOCUSS, L1LS and CoSaMP.

Two variants of the L0LS-CD algorithm was also developed in this chapter. V-L0LS-CD was developed for multivariate regression and gL0LS-CD can handle grouped variables. The simulation results show that similar to the performance of L0LS-CD, V-L0LS-CD initialized with the estimate of V-L1LS-CD produces

superior results when compared with existing algorithms. The performance of gL0LS-CD will be analyzed in the context of sparse network topology identification application in chapter 4.

2.A Appendix: FSEL, CLEAN and OMP

This section gives an overview of the widely used greedy algorithms FSEL, CLEAN and OMP. FSEL and OMP will be discussed first, followed by the CLEAN algorithm.

In FSEL and OMP estimate of β in each iteration is calculated as the least squares estimator over the current active set. Suppose k columns of X have already been chosen; collect these into a matrix X^k and denote the set of indices of the active set as Γ_c^k. let z be the new column being added and denote $X^{k+1} = [X^k, z]$. Denote $M = X^{k,T}X^k$, which is invertible for over-determined systems and for under-determined systems with $k < \text{rank}(X)$. The ordinary least squares estimator is then,

$$\beta^k = M^{-1}X^{k,T}y \Rightarrow M\beta^k = X^{k,T}y \tag{2.47}$$

Partition the updated least squares estimator as $\hat{\beta} = \binom{\beta^{k+1}}{b}$. Since $z^Tz = 1$ the normal equations are then

$$\begin{pmatrix} M & X^{k,T}z \\ z^TX^k & 1 \end{pmatrix} \begin{pmatrix} \beta^{k+1} \\ b \end{pmatrix} = \begin{pmatrix} X^{k,T}y \\ z^Ty \end{pmatrix} \tag{2.48}$$

Expanding the equations out, denoting $\rho = X^{k,T}z$ and using (2.47) gives

$$M\beta^{k+1} + \rho b = M\beta^k \tag{2.49}$$

$$\rho^T\beta^{k+1} + b = z^Ty \tag{2.50}$$

Using (2.49) we find,

$$\beta^{k+1} = \beta^k - M^{-1}\rho b \tag{2.51}$$

Then substituting (2.51) in (2.50) we get

$$\begin{aligned} \rho^T\beta^k + b(1-\rho^T M^{-1}\rho) &= z^T y \\ b &= (z^T y - \rho^T\beta^k)/d \end{aligned} \tag{2.52}$$

where $d = 1-\rho^T M^{-1}\rho$. But note that $z^T y - \rho^T\beta^k = z^T(y - X^k\beta^k) = z^T e^k$ so that $b = z^T e^k/d$. The error signal is,

$$\begin{aligned} e^{k+1} &= y - X^{k+1}\hat{\beta} \\ &= y - [X^k, z]\begin{pmatrix}\beta^{k+1}\\ b\end{pmatrix} \\ &= y - X^k\beta^{k+1} - z\rho \\ &= y - X^k(\beta^k - M^{-1}\rho b) - zb \\ &= y - X^k\beta^k - b(z - X^k M^{-1}\rho) \\ &= e^k - b(z - X^k M^{-1}\rho) \end{aligned} \tag{2.53}$$

The prior error signal energy is

$$\begin{aligned} \|e^k\|^2 &= \|y - X^k\beta^k\|^2 \\ &= \|y - [X^k, z]\begin{pmatrix}\beta^k\\ 0\end{pmatrix}\|^2 \\ &= \|y - X^{k+1}\hat{\beta} + X^{k+1}\left(\hat{\beta} - \begin{pmatrix}\beta^k\\ 0\end{pmatrix}\right)\|^2 \\ &= \|e^{k+1}\|^2 + \left(\hat{\beta} - \begin{pmatrix}\beta^k\\ 0\end{pmatrix}\right)^T X^{k+1,T} X^{k+1}\left(\hat{\beta} - \begin{pmatrix}\beta^k\\ 0\end{pmatrix}\right) \\ &= \|e^{k+1}\|^2 + \varepsilon \end{aligned}$$

However

$$\begin{aligned}
\varepsilon &= (\beta^{k+1,T} - \beta^{k,T}, b) X^{k+1,T} X^{k+1} \begin{pmatrix} \beta^{k+1} - \beta^k \\ b \end{pmatrix} \\
&= b^2(-\rho^T M^{-1}, 1) \begin{pmatrix} M & \rho \\ \rho^T & 1 \end{pmatrix} \begin{pmatrix} -M^{-1}\rho \\ 1 \end{pmatrix} \\
&= b^2(-\rho^T + \rho^T, -\rho^T M^{-1}\rho + 1) \begin{pmatrix} -M^{-1}\rho \\ 1 \end{pmatrix} \\
&= b^2(1 - \rho^T M^{-1}\rho) \\
&= b^2 d \\
&= (z^T e^k)^2/d
\end{aligned}$$

Thus

$$\|e^{k+1}\|^2 = \|e^k\|^2 - (z^T e^k)^2/d \tag{2.54}$$

2.A.1 FSEL

(i) From (2.54) we can see that the energy of the error signal is minimized when $(z^T e^k)^2/d$ is maximized. Thus given X^k, β^k and e^k, find

$\hat{u} = \arg.\max_{u \notin \Gamma_c^k} \dfrac{|x_{(u)}^T e^k|}{\Delta_u}$,

where $\Delta_u = \sqrt{1 - \rho_u^T (X^{k,T} X^k)^{-1} \rho_u} = \sqrt{d_u}$, $\rho_u = X^{k,T} x_{(u)}$.

Set $z = x_{(\hat{u})}, \gamma = z^T e^k, \rho = X^{k,T} z, \Delta = \Delta_{\hat{u}}$ and $d = \Delta^2$. NB. For initial step take $\beta^k = 0 \Rightarrow e^k = y$ and $\Delta_u = 1$.

(ii) Stop if the stopping criterion is met. If not continue.

(iii) Get b from (2.52), β^{k+1} from (2.51) and e^{k+1} from (2.53).

(iv) Update $e^k = e^{k+1}$, $\beta^k = \binom{\beta^{k+1}}{b}$ and return to (i)

2.A.2 OMP

OMP differs from FSEL only in the method of selecting the next index. So the OMP algorithm is the same as that presented in 2.A.1 with step (i) changed as

follows,

Given X^k, β^k and e^k, find $\hat{u} = \arg.\max_{u \notin \Gamma_c^k} |x_{(u)}^T e^k|$,

Set $z = x_{(\hat{u})}, \gamma = z^T e^k, \rho = X^{k,T} z, \Delta = \Delta_{\hat{u}}$ and $d = \Delta^2$.

2.A.3 CLEAN

Unlike FSEL and OMP clean algorithm does not orthogonally project the signal over the active set. At each iteration the β vector and the error signal is updated in a different way thus it does not follow the format given above. Initialize the algorithm with $\beta^0 = 0$.

(i) Given X, β^k and e^k, find $\hat{u} = \arg. \max_{u \notin \Gamma_c^{k-1}} |x_{(u)}^T e^k|$.

Set $z = x_{(\hat{u})}$ and $\gamma = z^T e^k$.

(ii) Update $\beta^{k+1} = \beta^k + \alpha\gamma\delta_{\hat{u}}$ and $e^{k+1} = e^k - \alpha\gamma z$ where α is a gain factor and $\delta_{\hat{u}}$ is a vector of 0s but with 1 in position $\hat{u}$ [220].

(iii) Stop if the stopping criterion is met. If not return to (i).

Generally the gain factor $\alpha = 1$ since it gives the fastest convergence. However as long as $0 < \alpha < 2$ the algorithm will converge [220]. Expression for the energy signal is as follows,

$$\|e^{k+1}\|^2 = \|e^k\|^2 - \alpha(2-\alpha)\gamma^2. \tag{2.55}$$

2.B Appendix: L0LS-CD Performance Comparison Continued

Section 2.11.5 gives results of a simulation that compares the performance of the L0LS-CD algorithm with other scalar regression algorithms. These simulations investigate how the performance of algorithms is affected by the distribution of the non-zero entries of the coefficient vector at various sparsity and noise levels. Due to the large volume of simulation results, some of the results are included in this appendix to improve the readability of section 2.11.5.

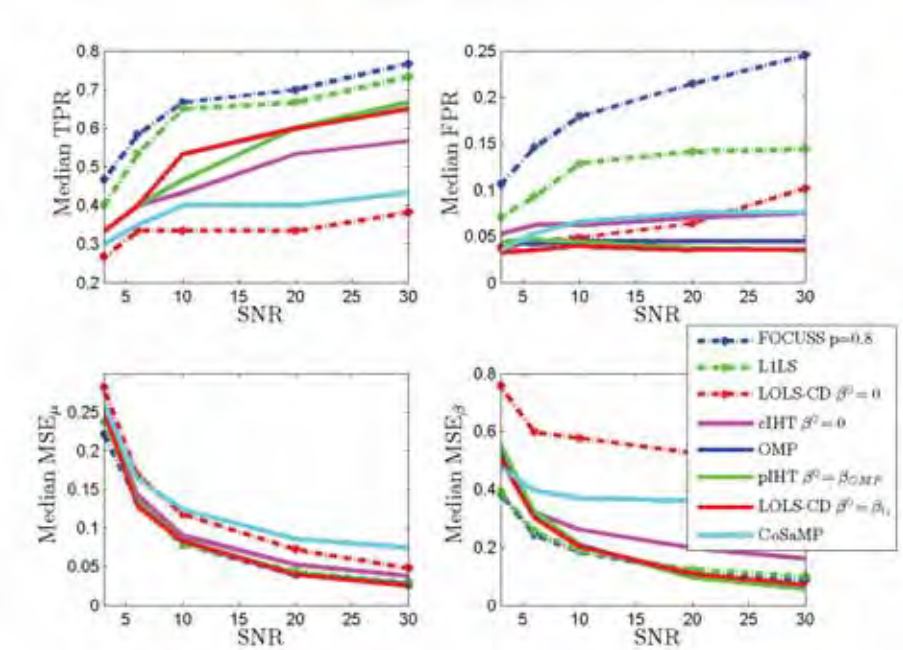

Figure 2.9: Performance comparison of L0LS-CD by the variation of the performance measures as a function of SNR in an under-determined system with n = 50, p = 128, r = 15 and non-zero values of $\beta^\star$ drawn from a Laplace distribution.

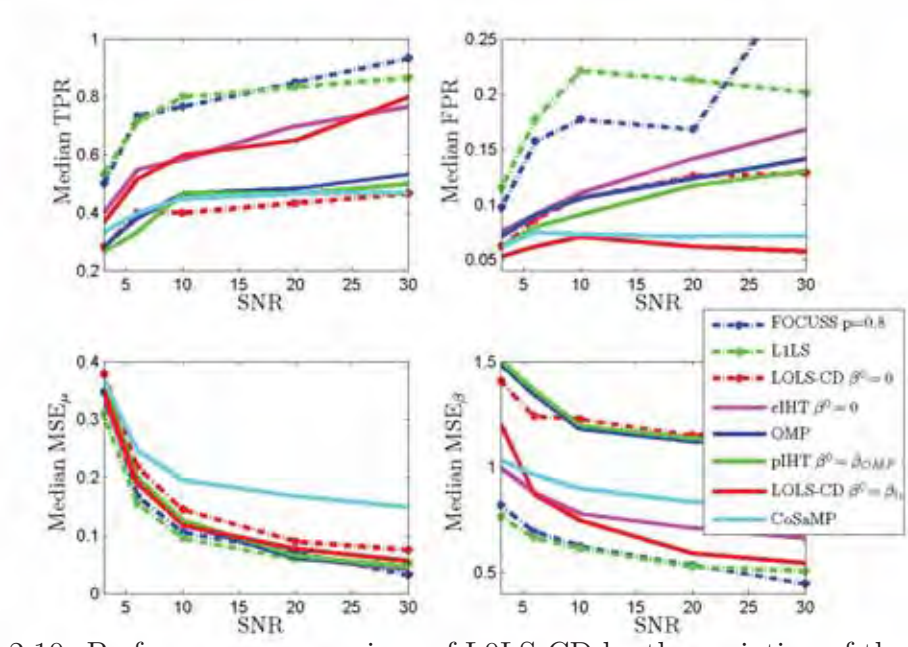

Figure 2.10: Performance comparison of L0LS-CD by the variation of the performance measures as a function of SNR in an under-determined system with n = 50, p = 128, r = 15 and non-zero values of $\beta^\star$ drawn from a Bernoulli distribution.

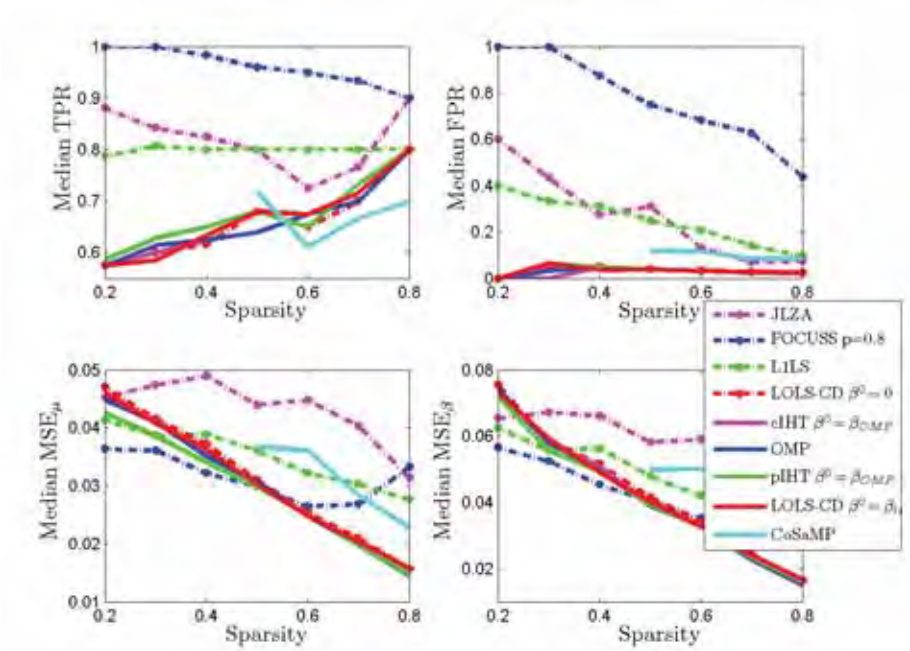

Figure 2.11: Performance comparison of L0LS-CD by the variation of the performance measures as a function of sparsity in an over-determined system with n = 128, p = 50, SNR = 10 and non-zero values of $\beta^\star$ drawn from a Laplace distribution.

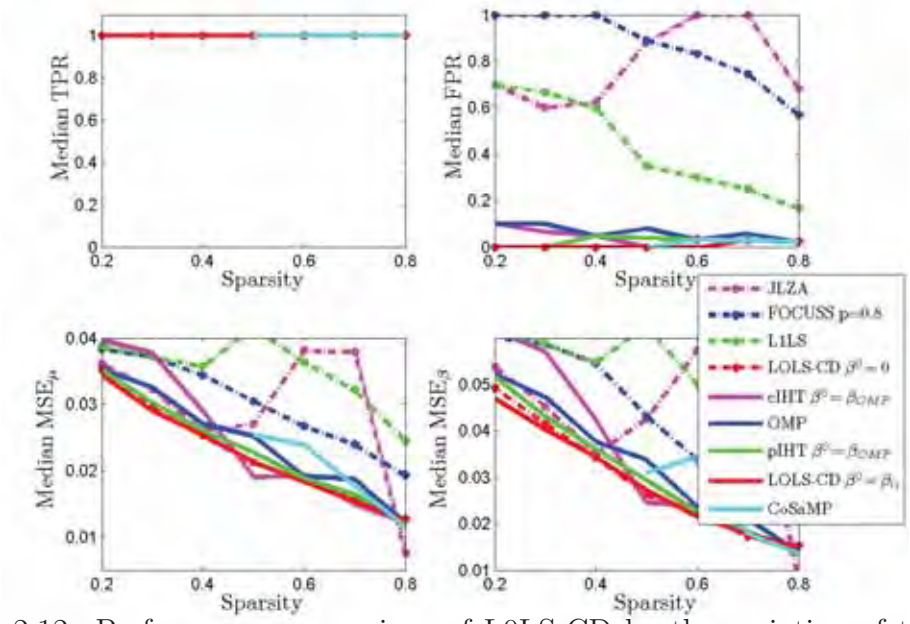

Figure 2.12: Performance comparison of L0LS-CD by the variation of the performance measures as a function of sparsity in an over-determined system with n = 128, p = 50, SNR = 10 and non-zero values of $\beta^\star$ drawn from a Bernoulli distribution.

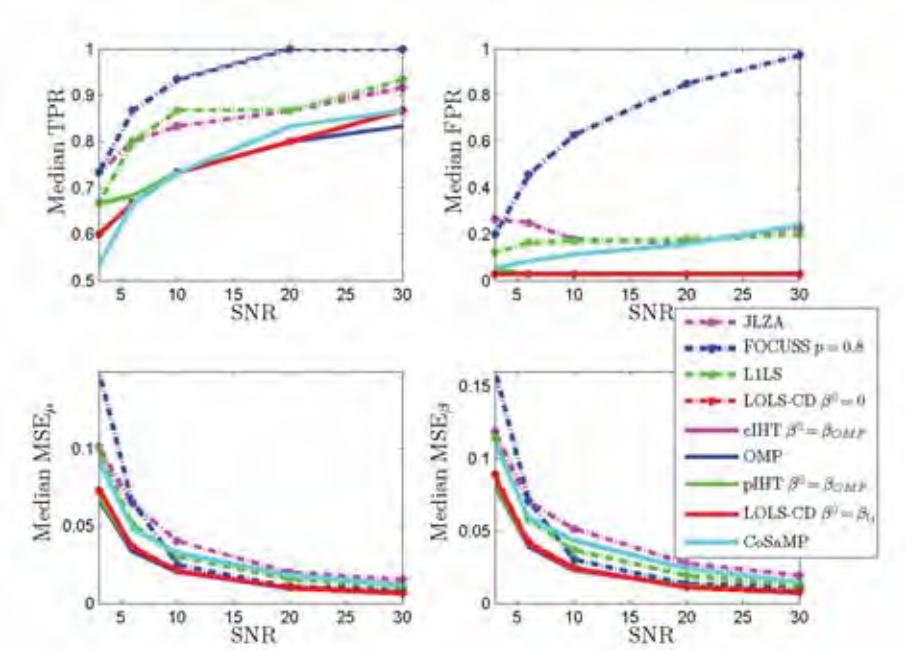

Figure 2.13: Performance comparison of L0LS-CD by the variation of the performance measures as a function of SNR in an over-determined system with n = 128, p = 50, r = 15 and non-zero values of $\beta^\star$ drawn from a Laplace distribution.

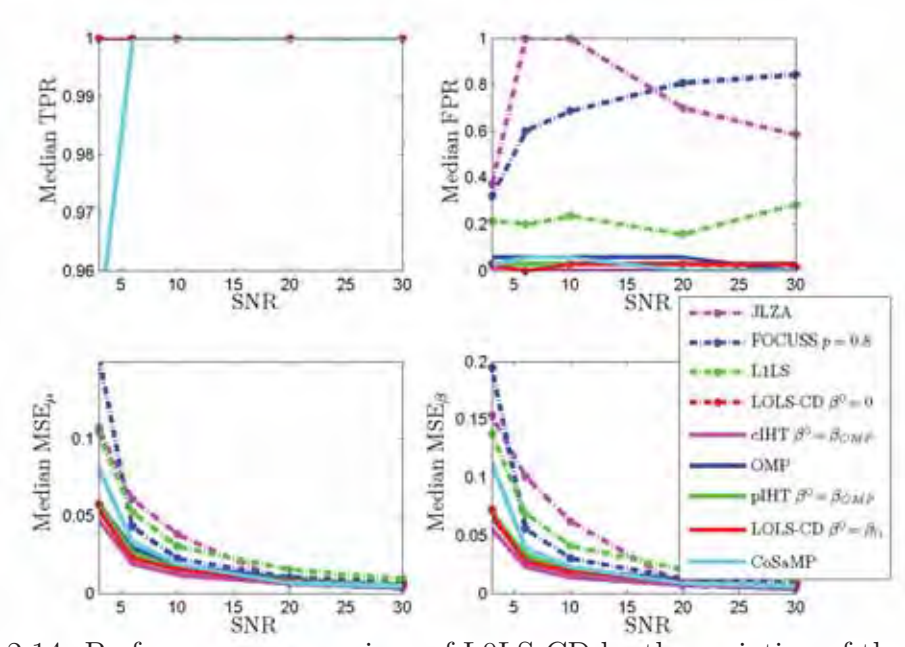

Figure 2.14: Performance comparison of L0LS-CD by the variation of the performance measures as a function of SNR in an over-determined system with n = 128, p = 50, r = 15 and non-zero values of $\beta^\star$ drawn from a Bernoulli distribution.

Chapter 3

Quadratic Concave Algorithm for Sparsity

The motivation behind sparse signal processing and its vast array of applications were outlined in the introduction chapter. As mentioned in section 1.5 optimizing the least squares criterion penalized with a non-quadratic penalty is one of the most successful methods of sparse modeling. Optimizing the l_0 penalized least squares (L0LS) criterion was addressed in chapter 2.

Due to the discrete non-convex nature of the l_0 norm, finding the global minimum of L0LS is NP hard. Thus the l_0 norm is commonly replaced by smoothed penalties as mentioned in section 1.5.3.3. Such smooth approximations are also available for other penalties such as the l_1 norm.

All of the algorithms discussed and developed in chapters 1 and 2 are concentrated on optimizing a criterion with a particular penalty. Few papers have developed generic algorithms that can handle a class of penalty functions. A family of non-convex penalties which can be decomposed as a difference of convex functions (DC) is considered in [76] which then uses DC programming [107] to optimize the criterion. The resulting iterative algorithm solves a convex weighted LASSO problem at each iteration. Since the negative of a convex function is concave, we could regard this as a problem of minimizing a sum of a convex and a concave function [279]. Drawback of this method is that many penalties cannot be decomposed as DC functions. Group of non-smooth, possibly non-convex

penalties are considered in [273] which develops a method based on proximity algorithms [39].

Majorization minimization (MM) algorithm was first introduced by [174] and then later developed by [127]. The acronym MM first appears in [112] and it has been used in many statistical applications [113]. [20] has used it to optimize L0LS and [150] has used MM technique to optimize an l_q penalized least squares criterion. This chapter develops an algorithm that can optimize least squares criterion penalized with a penalty that have what we call the quadratic concave property based on the MM technique. It is called the QC (quadratic concave) algorithm. The class of penalties QC supports is more general than that of [76, 273].

Section 3.1 introduces the class of quadratic concave penalties that will be considered in this chapter along with their properties. Section 3.2 gives the informal development of the algorithm followed by a comparison with a Newton algorithm. A formal development follows in section 3.3. Convergence analysis is provided in section 3.4. Simulations are in section 3.5 and conclusions in section 3.6.

3.1 Class of Quadratic Concave Penalties

Smooth approximations of sparse penalties has a long history including work in the image processing literature [77, 78, 131, 227] where the penalty is on a gradient rather than an amplitude as here. Denote $\rho(\beta)$ as the smooth approximating penalty. The algorithm developed in this chapter can handle any penalty $\rho(\beta)$ which adhere to the following properties.

(i) $\rho(\beta)$ is differentiable.

(ii) $\rho(\beta)$ is quadratic concave i.e. $\rho(\beta) = \kappa(\beta^2)$ where $\kappa(\cdot)$ is concave.

Examples of smooth approximating penalties $\rho(\cdot)$ and their properties are collected in table 3.1. In the sequel a fundamental role is played by the weight function $\omega(\beta) = \rho'(\beta)/\beta$ so these are also listed.

Table 3.1: Smooth Approximations to Sparse Penalties and their Properties.

Penalty	Approximation $\rho(\cdot)$	$\rho'(\beta)$	$\omega(\cdot)$	$\rho''(\beta)$	Source
l_1	$\sqrt{\beta^2+\gamma^2}-\gamma$	$\frac{\beta}{\sqrt{\beta^2+\gamma^2}}$	$\frac{1}{\sqrt{\beta^2+\gamma^2}}$	$\frac{1}{\sqrt{\beta^2+\gamma^2}}-\frac{\beta^2}{[\beta^2+\gamma^2]^{3/2}}$	[259]
l_1	$\gamma \ln\cosh(\frac{\beta}{\gamma})$	$\tanh(\frac{\beta}{\gamma})$	$\frac{\tanh(\frac{\beta}{\gamma})}{\beta}$	$\frac{1}{\gamma}(1-\tanh^2(\frac{\beta}{\gamma}))$	[199]
l_0	$1-e^{-\frac{\beta^2}{2\gamma^2}}$	$\frac{\beta}{\gamma^2}e^{-\frac{\beta^2}{2\gamma^2}}$	$\frac{1}{\gamma^2}e^{-\frac{\beta^2}{2\gamma^2}}$	$\frac{1}{\gamma^2}e^{-\frac{\beta^2}{2\gamma^2}}-\frac{\beta^2}{\gamma^2}e^{-\frac{\beta^2}{2\gamma^2}}$	[159]
-	$\log(\sqrt{\beta^2+\gamma^2})-\log(\gamma)$	$\frac{\beta}{\beta^2+\gamma^2}$	$\frac{1}{\beta^2+\gamma^2}$	-	[269, 270]

Note that the first two entries of table 3.1 have non-decreasing $\rho'(\beta)$ and are thus convex. The third and fourth entries are neither convex nor concave. The algorithm developed below can handle all these cases.

The fifth column of table 3.1 gives the second derivative of $\rho(\beta)$. When $\rho''(\beta)$ exists, it has the form $\omega(\beta)-\beta^2\psi(\beta)$ and $\omega(\cdot)$, $\psi(\cdot)$ are both non-negative. This feature is crucial in our ensuing informal discussion of convergence given in section 3.2. The formal convergence analysis does not require the existence of a second derivative.

When the second derivative exists this structure is general as follows. Inspection of the smoothed approximations in the table 3.1 yields the following crucial observation,

$$\rho(\beta) = \kappa(\beta^2), \tag{3.1}$$

where $\kappa(x)$ is concave, thus $\rho(\beta)$ is quadratic concave. Then $\rho'(\beta) = 2\beta\kappa'(\beta^2)$ so that $\omega(\beta) = 2\kappa'(\beta^2)$ which by inspection is greater than 0 in each case. We can now write $\rho'(\beta) = \beta\omega(\beta)$. Then

$$\begin{aligned}\rho''(\beta) =& \omega(\beta) + \beta\omega'(\beta) = \omega(\beta) + 4\beta^2\kappa''(\beta^2),\\ =& \omega(\beta) - \beta^2\psi(\beta),\end{aligned} \tag{3.2}$$

where,

$$\psi(\beta) = -4\kappa''(\beta^2) \geq 0, \tag{3.3}$$

since $\kappa(x)$ is concave.

This chapter develops an algorithm to optimize least squares criterion penalized by a quadratic concave penalty. Thus the criterion is given by (1.4c) with $f(\beta) = \rho(\beta)$,

$$J(\beta) = \|y - X\beta\|^2 + h\rho(\beta). \tag{3.4}$$

[180, 179] have considered a class of functions which has the same structure as the quadratic concave penalties discussed above. However [180, 179] do not consider a penalized procedure and do not have the algorithm developed below. Reweighted l_2 penalty discussed in [270] is a generalization of the log function given in table 3.1, however there are no convergence theorems; just an informal discussion of convergence. The reweighted l_1 penalty discussed in [28, 270] is not smooth and thus falls outside of the framework discussed above.

3.2 Informal Development of the QC Algorithm

This section presents a heuristic development of the algorithm assuming $\rho(\beta)$ is twice differentiable. In particular it will be compared with a Newton algorithm to emphasis its superiority.

3.2.1 Derivation

The first order optimality condition or Euler equation of (3.4) is,

$$\frac{dJ}{d\beta} = 0 = -X^T(y - X\beta) + h\text{diag}(\omega(\beta_u))\beta. \tag{3.5}$$

The form of this equation immediately suggests a fixed point iteration,

$$\beta^{k+1} = M_k^{-1} X^T y, \tag{3.6}$$

where

$$M_k = X^T X + hW_k, \tag{3.7}$$

and $W_k = \text{diag}(\omega(\beta_u^k))$. Note that M_k is positive definite.

3.2.2 Informal Analysis

If (3.4) is differentiated again and use the result in (3.2) we obtain,

$$\frac{d^2 J}{d\beta d\beta^T} = X^T X + h(W_k - Q_k), \tag{3.8}$$

where $Q_k = \text{diag}((\beta_u^k)^2 \psi(\beta_u^k))$ and is positive semi-definite. Now consider the Taylor series,

$$J(\beta^{k+1}) = J(\beta^k) + \frac{\partial J}{\partial \beta^k}^T \Delta_k + \frac{1}{2}\Delta_k^T \frac{d^2 J}{d\beta^k d\beta^{k,T}} \Delta_k + o(\|\beta^k\|^2), \tag{3.9}$$

where $\Delta_k = \beta^{k+1} - \beta^k$. Using (3.5) and (3.6),

$$\begin{aligned} \frac{dJ}{d\beta^k} &= (X^T X + hW_k)\beta^k - X^T y \\ &= -(X^T X + hW_k)(\beta^{k+1} - \beta^k) \\ &= -M_k \Delta_k. \end{aligned} \tag{3.10}$$

Now denote $J^{k+1} = J(\beta^{k+1}), J^k = J(\beta^k)$ and put (3.8) and (3.10) in (3.9) to yield,

$$\begin{aligned} J^{k+1} - J^k &= -\Delta_k^T M_k \Delta_k + \frac{1}{2}\Delta_k^T (M_k - Q_k)\Delta_k, \\ &= -\frac{1}{2}\Delta_k^T (M_k + Q_k)\Delta_k. \end{aligned} \tag{3.11}$$

so that we get, locally, a guaranteed reduction in $J(\cdot)$ unless $\Delta_k = 0$. It is very interesting to compare this with the Newton algorithm which using (3.7) and

(3.8) clearly has the form,

$$\begin{aligned}\beta^{k+1} &= \beta^k - (M_k - Q_k)^{-1}\frac{dJ}{d\beta^k} \\ &= \beta^k - (M_k - Q_k)^{-1}(M_k\beta^k - X^T y) \\ &= (M_k - Q_k)^{-1}[X^T y - Q_k\beta^k] \end{aligned} \tag{3.12}$$

The corresponding change in the criterion is then,

$$J^{k+1} - J^k = -\frac{1}{2}\Delta_k^T(M_k - Q_k)\Delta_k. \tag{3.13}$$

We now see that the Newton algorithm is inferior in two ways. Firstly since $M_k - Q_k$ is not guaranteed to be positive definite, the Newton algorithm may stall. Secondly even if it does not, the update is not guaranteed to reduce the value of the criterion. We have a remarkable situation where an MM algorithm is superior to the Newton algorithm.

3.3 Formal Development of QC Algorithm

Section 3.3.1 will review briefly the idea of MM algorithms and then section 3.3.2 will develop the QC algorithm formally. Unlike the informal analysis, here $\rho(\beta)$ is only required to be differentiable.

3.3.1 MM Algorithms

The MM principle, which goes back to [174], is a principle for deriving iterative algorithms to minimize (or maximize) a criterion of interest. An MM algorithm is guaranteed to not increase (decrease) the criterion at each step. A very readable survey is available in [113] and more recently [275].

The idea is the following [113]. Let $J(\beta)$ be the criterion of interest and let β^k be the estimate at the k$^{\text{th}}$ iterate. Suppose we can find a bivariate functional $M(\beta|\beta^k)$ which obeys the following two properties.

(i) $M(\beta|\beta^k) \geq J(\beta)$ for all β ($M(\beta|\beta^k)$ Majorizes $J(\beta)$).

(ii) $M(\beta^k|\beta^k) = J(\beta^k)$.

Thus $M(\beta|\beta^k)$ lies above $J(\beta)$ and is tangent at β^k. Then we generate iterates as follows,

$$\beta^{k+1} = \arg.\min_{\beta} M(\beta|\beta^k). \tag{3.14}$$

We then find,

$$J(\beta^{k+1}) = M(\beta^{k+1}|\beta^k) + [J(\beta^{k+1}) - M(\beta^{k+1}|\beta^k)]. \tag{3.15}$$

But from (3.14), the first term is less than $M(\beta^k|\beta^k)$ while the second is less than 0. We thus find,

$$J(\beta^{k+1}) \leq M(\beta^k|\beta^k) = J(\beta^k). \tag{3.16}$$

which demonstrates the required non-increase.

The question of course is how to construct the majorizing functional. Unlike the EM algorithm there is no principled way to do this. Rather there is a growing body of methods and [113] gives some of them and one of those is relevant here. Consider the Taylor series,

$$J(\beta) = J(\beta^k) + \frac{\partial J}{\partial \beta^k}^T (\beta - \beta^k) + \frac{1}{2}(\beta - \beta^k)^T \frac{\partial^2 J}{\partial \beta_* \partial \beta_*}(\beta - \beta^k), \tag{3.17}$$

where β_* is an intermediate value between β^k, β^{k+1}. Now suppose we can find a bound H on the Hessian such that $H - \frac{\partial^2 J}{\partial \beta \partial \beta}$ is positive semi-definite for all β. Then we can take

$$M(\beta|\beta^k) = J(\beta^k) + \frac{\partial J}{\partial \beta^k}^T (\beta - \beta^k) + \frac{1}{2}(\beta - \beta^k)^T H(\beta - \beta^k), \tag{3.18}$$

And indeed we have $M(\beta|\beta^k) \geq J(\beta)$ for all β. As well as $M(\beta^k|\beta^k) = J(\beta^k)$. Thus the QC iterate will be

$$\beta^{k+1} = \beta^k + H^{-1} \frac{\partial J}{\partial \beta^k}. \tag{3.19}$$

Using this result appendix 3.A shows that proximity algorithms of [39] are MM algorithms. Also [225] has recognized the concave-convex (CC) procedure as an MM algorithm, and a proof is given in appendix 3.B for completeness.

But it turns out in problem (3.4) such a β free bound cannot be found. Rather a β dependent bound $W(\beta)$ is available. But drawing on the more sophisticated argument of [109] and the additive nature of (3.4) criterion, it is possible nevertheless to construct an MM algorithm.

3.3.2 QC Algorithm: Formal Development

Now it is formally shown that the algorithm developed in section 3.2 is indeed an MM algorithm. First introduce the condition,

Condition 1: $\omega(\theta)$ is decreasing in θ.

Lemma L7. Since $\rho(\beta) = \kappa(\beta^2)$, $\omega(\theta)$ is decreasing iff $\kappa(|\theta|)$ is strictly concave.

Proof. $\rho'(\beta) = 2\beta\kappa'(\beta^2)$. Thus $\omega(\beta) = 2\kappa'(\beta^2)$. And so $\omega(\beta)$ is decreasing iff $\kappa(|\beta|)$ is strictly concave.

When $\rho(\cdot)$ is twice differentiable we already saw in (3.2) that concavity of $\kappa(\cdot)$ is equivalent to $\psi(\cdot) \geq 0$. Crucial to our discussion is the following lemma inspired by the method in [109].

Lemma L8. let θ be a scalar. Given a reference value $\hat{\theta}$ introduce the function,

$$V(\theta|\hat{\theta}) = \rho(\hat{\theta}) + \frac{1}{2}\omega(\hat{\theta})(\theta^2 - \hat{\theta}^2). \tag{3.20}$$

Clearly $V(\hat{\theta}|\hat{\theta}) = \rho(\hat{\theta})$. Further, provided condition 1 holds, $V(\theta|\hat{\theta}) \geq \rho(\theta)$ for all θ.

Proof. Put $d(\theta) = V(\theta|\hat{\theta}) - \rho(\hat{\theta})$. Then $d(-\theta) = d(\theta)$ and $d(\hat{\theta}) = 0$. We have to show $d(\theta) \geq 0$. Next

$$d'(\theta) = \theta(\omega(\hat{\theta}) - \omega(\theta)). \tag{3.21}$$

Since $d(\theta)$ is an even function we now need only consider $\theta > 0$. There are two cases.

If $0 < \hat{\theta} < \theta$ then $\omega(\hat{\theta}) > \omega(\theta) \Rightarrow d'(\theta) \geq 0$. Integrating this gives $d(\theta) - d(\hat{\theta}) \geq 0 \Rightarrow d(\theta) \geq 0$.

If $0 < \theta < \hat{\theta}$ then $\omega(\hat{\theta}) < \omega(\theta) \Rightarrow d'(\theta) \leq 0$. Integrating this gives $d(\hat{\theta}) - d(\theta) \leq 0 \Rightarrow d(\theta) \geq 0$. The proof is complete.

This result is used to construct a majorizing function for (3.4). Indeed using the lemma L8 consider that

$$\begin{aligned} M(\beta|\beta^k) &= \frac{1}{2}\|y - X\beta\|^2 + h\sum_1^p V(\beta_u|\beta_u^k) \\ &\geq \frac{1}{2}\|y - X\beta\|^2 + h\sum_1^p \rho(\beta_u) = J(\beta). \\ M(\beta^k|\beta^k) &= \frac{1}{2}\|y - X\beta^k\|^2 + h\sum_1^p \rho(\beta_u^k) = J(\beta^k) \end{aligned}$$

So $M(\beta|\beta^k)$ is a majorizing function. Further setting $\frac{d}{d\beta}M(\beta|\beta^k) = 0$ yields

$$\begin{aligned} 0 &= -X^T(y - X\beta) + hW_k\beta \\ \Rightarrow \beta &= \beta^{k+1} = (X^TX + hW_k)^{-1}X^Ty, \end{aligned} \tag{3.22}$$

which is exactly our earlier update (3.6) and exhibits it as an MM update. Thus $J(\beta^{k+1}) \leq J(\beta^k)$. Similar to the L0LS-CD algorithm, QC can be terminated when $J(\beta^k) - J(\beta^{k+1}) \leq$ tolerance. Initialization of QC with respect to smooth approximations of the l_0 penalty is discussed in section 3.5.

3.4 Convergence

Although the development of the QC algorithm supports both over-determined and under-determined systems, the stability analysis given here applies only to over-determined systems. The QC algorithm can be successfully applied to under-determined systems as shown by simulation results in section 3.5. Thus the

stability analysis of QC in under-determined systems is open for future research.

Note that $M(\beta|\beta^k)$ is quadratic in β and so obeys an exact second order Taylor series. We have,

$$\begin{aligned} M(\beta|\beta^k) &= \frac{1}{2}\|y - X\beta\|^2 + h\sum_1^p \rho(\beta_u^k) + \frac{h}{2}[\beta^T W_k \beta - \beta^{k,T} W_k \beta^k] \\ \Rightarrow \frac{dM(\beta|\beta^k)}{d\beta} &= -X^T(y - X\beta) + hW_k\beta \\ &= M_k\beta - X^T y \\ \Rightarrow \frac{d^2M(\beta|\beta^k)}{d\beta d\beta^T} &= M_k = X^T X + hW_k \end{aligned}$$

Thus since $M(\beta^{k+1}|\beta^k) = J^{k+1}$ and $M(\beta^k|\beta^k) = J^k$ we find,

$$\begin{aligned} J^{k+1} &= J^k - \Delta_k^T M_k(\beta^{k+1} - \beta^k) + \frac{1}{2}\Delta_k^T M_k \Delta_k, \\ &= J^k - \frac{1}{2}\Delta_k^T M_k \Delta_k, \\ &\leq J^k - \frac{1}{2}\Delta_k^T X^T X \Delta_k, \end{aligned}$$

Introduce,

Condition 2: $X^T X$ has full rank.

Result I. Under conditions 1 and 2 as $m \to \infty$,

(a) J^k converges to a limit $J_\infty \geq 0$.

Proof. We have $0 \leq J(\beta^k) \leq J(\beta^{(0)})$. Thus J^k is a bounded sequence which is non-increasing and so must have a limit say J_∞. Thus $J^{k+1} - J^k \to J_\infty - J_\infty = 0$.

(b) $\Delta_k = \beta^{k+1} - \beta^k \to 0$.

Proof. Let σ be the smallest eigenvalue of $X^T X$ then,

$\frac{1}{2}\sigma\|\Delta_k\|^2 \leq \frac{1}{2}\Delta_k^T X^T X \Delta_k = J^{k+1} - J^k \to 0$. Thus $\Delta_k \to 0$.

(c) The limit points of the β^k sequence form a compact connected set.

Proof. In view of (b) this follows from Ostrowski's theorem [176].

(d) The limit points are stationary points of $J(\beta)$.

Proof. In view of (b) any limit point β_* must obey,

$\beta_* = (X^T X + hW(\beta_*))^{-1} X^T y \Rightarrow \frac{dJ}{d\beta}|_{\beta_*} = 0.$

Essentially the same algorithm is developed in [111] and is also recognized as being an MM algorithm. However there is a fundamental and significant difference between [111] and the present results. QC algorithm is derived and analyzed under conditions 1 and 2, whereas [111] have the extremely stringent requirement that $\rho(\cdot)$ be concave; this would rule out three of our four cases listed in table 3.1. Thus under the results of [111] the QC algorithm can only be used in very restrictive circumstances. Whereas our results show it to be of very wide applicability.

3.5 QC Simulation

Although smooth approximations of the l_1 penalty are supported by the QC algorithm, only the approximations of the l_0 penalty will be considered in this section. Performance of the QC algorithm will be analyzed based on the G_0^γ penalty (1.2f) which will be called QC-$G_{0,\gamma}$ and the $TH_0^{\gamma,b}$ penalty (1.2e) with $b = 2$ which will be called QC-$TH_{0,\gamma}$.

As shown in section 2.11.5 most algorithms performed well on over-determined systems, thus only under-determined systems are considered here. However it should be noted that the QC algorithm can be applied to over-determined systems. Simulations were set up in a similar way as that of L0LS-CD performance comparison simulations given in section 2.11. X, y and β were generated in the same way with SNR $= 10$, $n = 50$, $p = 128$ and the same performance measures were used.

L0LS-CD and the other algorithms considered in section 2.11 required the selection of h. Smooth approximation penalties require selection of at least one more tuning parameter, e.g. γ for the G_0^γ and $TH_0^{\gamma,b}$ penalties. Thus unlike the algorithms considered in section 2.11, QC algorithm require selection of two or more tuning parameters. Here the tuning parameters are selected using BIC. The

BIC criterion is the same as that given in section 2.11.2, however it is minimized with respect to two tuning parameters as opposed to just h in section 2.11.

Although the smooth approximations of the l_0 norm have the advantage of being differentiable they still have the drawback of having multiple local minima. Thus similar to L0LS-CD, proper initialization is essential for the proper functioning of the QC algorithm with l_0 approximations. Thus a method similar to that used in section 2.11.3 was employed. QC algorithm was initialized with all zeros $\beta^0 = 0$, output of L1LS $\beta^0 = \beta_{l_1}$ and output of OMP $\beta^0 = \beta_{OMP}$ to find out the best method of initialization.

Many sparse approximating algorithms were compared with the L0LS-CD algorithm in section 2.11.5 and L0LS-CD outperformed all of them. Thus the performance of those algorithms will not be re-evaluated here. Since the smooth approximations of the l_0 norm is considered the performance of the QC algorithm is compared with the l_0 denoising algorithms L0LS-CD and pIHT. Scalar version of JLZA algorithm is not considered here due to the issues mentioned in section 2.11.4.

Sparsity of the β vector was varied from 0.5 to 1 and 20 β vectors were generated for each sparsity level. 20 y vectors were generated for each β vector. Performance of QC-$G_{0,\gamma}$ is given in figure 3.1 and that of the QC-$TH_{0,\gamma}$ is given in figure 3.2.

It is clear from figures 3.1 and 3.2 that proper initialization is vital for both the G_0^γ and $TH_0^{\gamma,b}$ penalties. Similar to the observation made in section 2.11.5 QC algorithm with both penalties seem to perform best when initialized with the output of L1LS. Both penalties with $\beta^0 = \beta_{l_1}$ have a higher FPR than L0LS-CD. However at very low sparsity levels the TPR of L0LS-CD and pIHT drop dramatically but QC algorithm with smooth approximations of the l_0 penalty manage to maintain comparatively higher TPR and therefore has lower MSE_μ. Thus in this simulation, the QC algorithm with smooth approximations of the l_0 penalty outperform L0LS-CD only at low sparsity levels.

In section 2.11 and in the example given above the X matrix was generated from independent Gaussians with zero mean and unit variance. The profile of the singular values (SV) of such a matrix is given in figure 3.3(a). The SV profiles of regression matrices in real world applications may not be so well behaved.

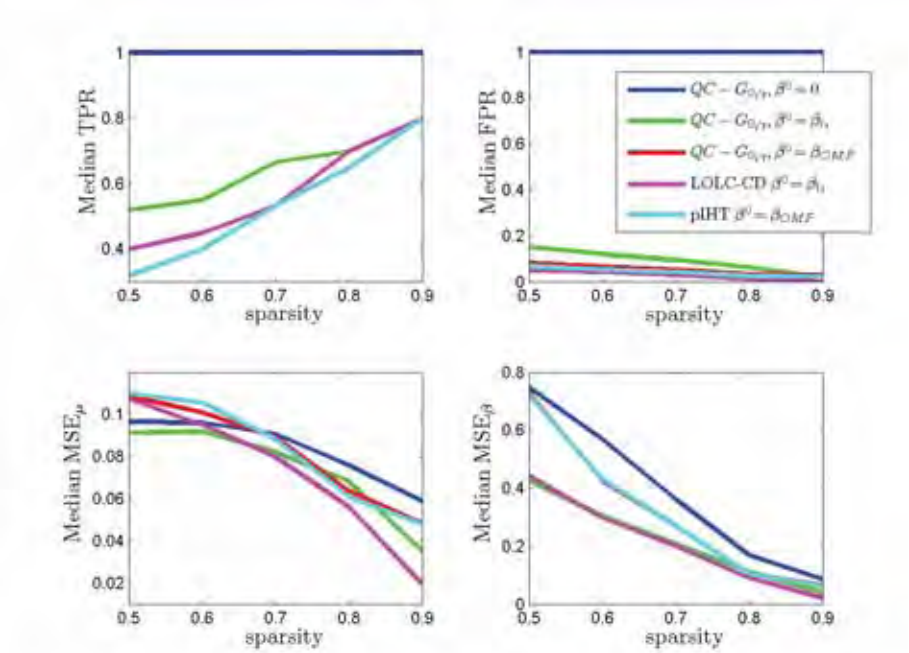

Figure 3.1: Performance comparison of QC-$G_{0,\gamma}$ by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128 and SNR = 10.

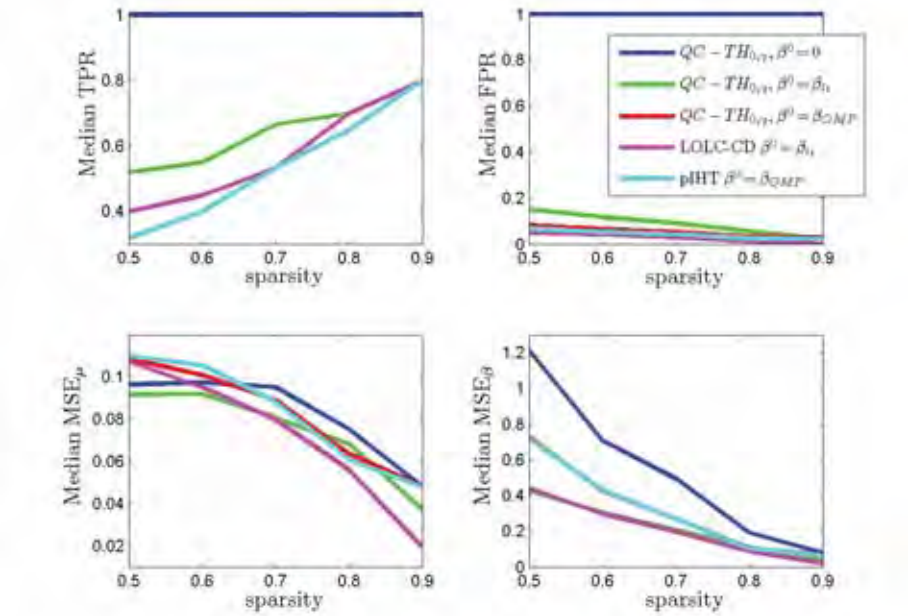

Figure 3.2: Performance comparison of QC-$TH_{0,\gamma}$ by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128 and SNR = 10.

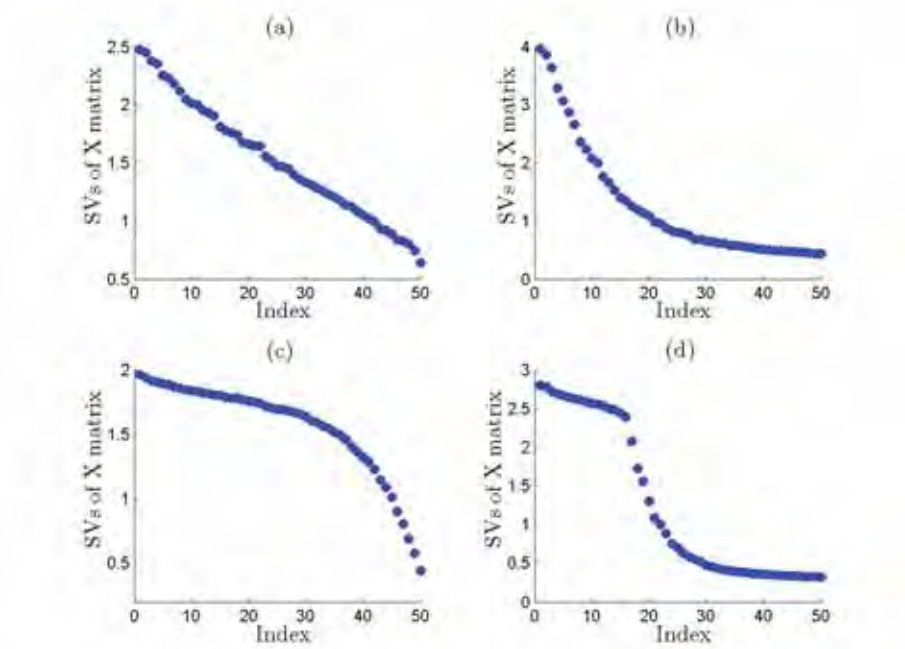

Figure 3.3: Singular value profiles of X matrices.

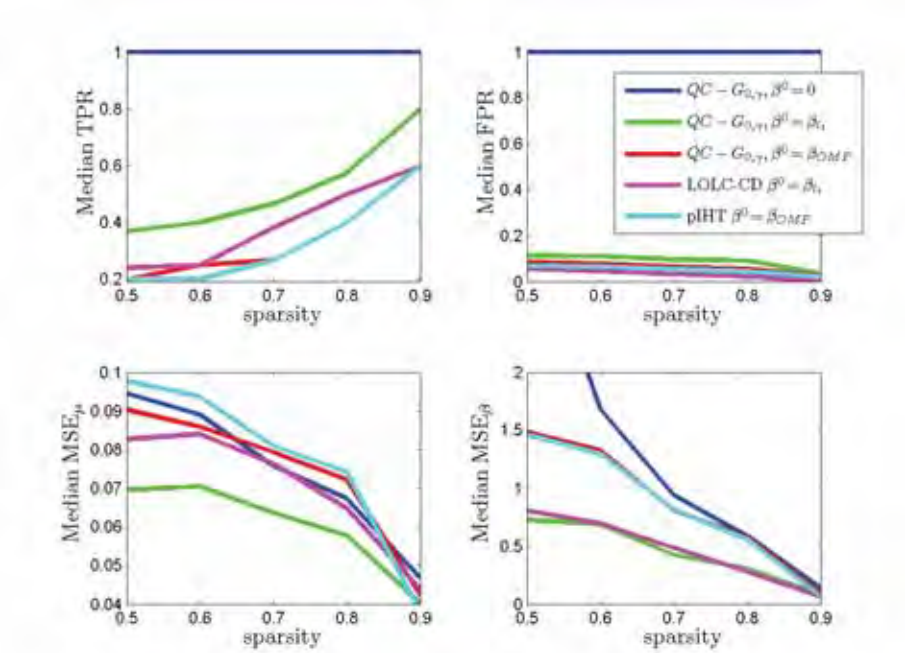

Figure 3.4: Performance comparison of QC-$G_{0,\gamma}$ by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128, SNR = 10 and when X has singular value profile (d).

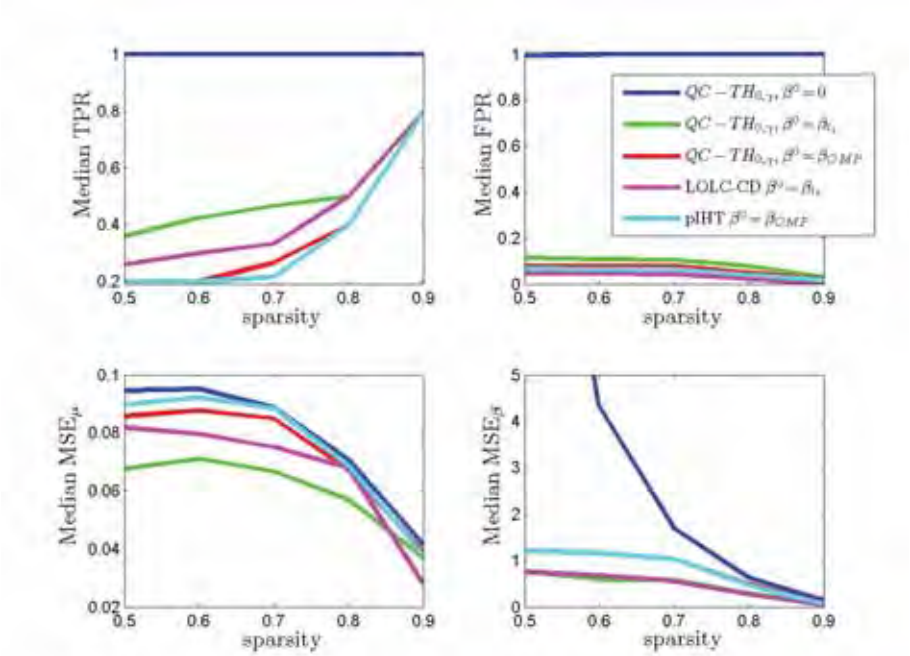

Figure 3.5: Performance comparison of QC-$TH_{0,\gamma}$ by the variation of the performance measures as a function of sparsity in an under-determined system with n = 50, p = 128, SNR = 10 and when X has singular value profile (d).

Thus X matrices with three different SV profiles as given by figures 3.3(b), 3.3(c) and 3.3(d) were generated and their columns of were scaled to have unit norm. Simulation described above was repeated with the newly generated X matrices. Profile 3.3(d) can be considered as an intermediate between profiles 3.3(b) and 3.3(c) as it has a range of SV with high values and a range with values close to zero. Thus only the results of the simulation done with profile 3.3(d) will be shown here. Performance of the QC algorithm on a system where the SV profile of the X matrix is similar to that shown in 3.3(d) is given in figures 3.4 and 3.5.

Its clear from figures 3.4 and 3.5 that the performance of all the algorithms degrade when many SVs of X are close to zero. However compared to L0LS-CD and pIHT, QC algorithm with $\beta^0 = \beta_{l_1}$ has much higher TPR while maintaining low FPR and as a result its has the lowest MSE_μ and MSE_β. Its MSE_μ in particular is much lower than the L0LS-CD and pIHT. Furthermore QC-$G_{0,\gamma}$ seem to perform better than QC-$TH_{0,\gamma}$. From these simulations it is clear that while L0LS-CD is the best option for sparse estimation when few SVs of X are close to zero, when this condition is not met QC algorithm with smooth approximations

of the l_0 norm is a better option.

3.6 Conclusions

This chapter developed an algorithm called QC, based on MM technique that can optimize the least squares criterion penalized with a quadratic concave penalty. Informal development of the QC algorithm showed that it locally reduces the criterion at each iteration where as the corresponding Newton algorithm may not. A novel MM functional is used and convergence and the stability of the algorithm was also investigated. Simulation results show that when many singular values of the X matrix is close to zero the algorithm developed in this chapter outperforms L0LS-CD algorithm developed in chapter 2.

3.A Appendix: Proximity Algorithm as an MM Algorithm

The proximity algorithm [39] aims to solve the following optimization problem: $\min_x J(x) = L(x) + R(x)$ where $L(x)$ and $R(x)$ are convex. Also $R(x)$ is differentiable with Lipschitz continuous gradient with Lipschitz constant λ i.e.

$$||\nabla R(x) - \nabla R(y)|| \leq \lambda ||x - y||$$

Now by the mean value theorem [9][Theorem 12.9] there exists $\bar{y}$ lying between x, y i.e. $\bar{y} = tx + (1-t)y$ for some $0 < t < 1$ such that,

$$R(x) = R(y) + (x - y)^T \nabla R(\bar{y})$$

Adding and subtracting $(x - y)^T \nabla R(y)$ we find,

$$R(x) = R(y) + (x - y)^T \nabla R(y) + (x - y)^T [\nabla R(\bar{y}) - \nabla R(y)]$$

But convexity of $R(\cdot)$ ensures that for all x, y we have $R(x) > R(y) + (x -$

$y)^T \nabla R(y)$. We thus conclude that,

$$(x-y)^T[\nabla R(\bar{y}) - \nabla R(y)] \geq 0$$

We can then conclude

$$\begin{aligned} 0 &\leq (x-y)^T[\nabla R(\bar{y}) - \nabla R(y)] \\ &\leq ||x-y||||\nabla R(\bar{y}) - \nabla R(y)|| \\ &\leq ||x-y||\lambda||\bar{y}-y|| \leq \lambda||x-y||^2 \end{aligned}$$

Introducing the function

$$R(x,y) = R(y) + (x-y)^T \nabla R(y) + \lambda||x-y||^2$$

We thus conclude that $R(x,x) = R(x)$ while $R(x) \leq R(x,y)$. It follows that $M(x,y) = L(x) + R(x,y)$ forms a majorization function for $J(x)$. Thus we can generate a sequence of MM iterates according to,

$$\begin{aligned} x^{k+1} &= \arg\,.\min_x M(x, x^k) \\ &= \arg\,.\min_x L(x) + R(x^k) + (x-x^k)^T \nabla R(x^k) + \lambda||x-x^k||^2 \\ &= \arg\,.\min_x \frac{1}{\lambda} L(x) + ||x - x^k + \frac{1}{2\lambda}\nabla R(x^k)||^2 \\ &= \text{prox}_{\frac{1}{\lambda}L}(x^k - \frac{1}{2\lambda}\nabla R(x^k)) \end{aligned} \tag{3.23}$$

where

$$\text{prox}_{\frac{1}{\lambda}L}(y) = \arg\,.\min_x \frac{1}{\lambda} L(x) + ||x-y||^2$$

is exactly the proximity opertor defined in [39]. And this MM algorithm is then exactly the backward-forward splitting algorithm of [39][1.17] as claimed. Further we now see that the proximity operator appears naturally as a result of the quadratic majorization.

It also follows that the algorithms cited in [39] as examples of proximity

algorithms are MM algorithms.

3.B Appendix: CC as an MM Algorithm

For the DC problem we have $J(x) = L(x) - R(x)$ where $L(x)$, $R(x)$ are convex and $R(x)$ is differentiable. Thus as before $R(x) \geq R(y) + (x-y)^T \nabla R(y)$ and so as [225] observed,

$$M(x,y) = L(x) - R(y) - (x-y)^T \nabla R(y)$$

is a majorizer for $J(x)$. Since indeed $M(x,y) \geq J(x)$ while $M(x,x) = J(x)$. The MM iteration is then,

$$x^{k+1} = \arg\,.\min_x L(x) - x^T \nabla R(x^k)$$

This may now be compared with (3.23) the difference being the quadratic term due to the additive convexity as opposed to the subtracted convexity here.

Chapter 4

Application: Sparse Coloured System Identification

There has recently been growing interest in sparse transfer function estimation [22, 214, 189, 190]. Such approaches could prove useful in hardware implementation by saving on elements corresponding to zeroed coefficients. Furthermore enforcing sparsity on the estimates will reduce the variance and as a result increase the reliability of the estimates. Estimating a sparse transfer function can be posed as a problem of finding a sparse solution to a linear model such as (1.1). An overview of existing sparse approximating algorithms is given in section 1.5 and new algorithms have been proposed in chapters 2 and 3.

The traditional basis for transfer function model expansion is the delay (shift) operator. However due to its short memory, it is not suitable for systems with rapid sampling, as this leads to a substantial increase in the order of the approximated model [154, 264]. Recent work [140, 193, 198] has tried to overcome these issues by applying sparse regression based methods such as non-negative garrotte and LASSO, but unless one uses finite impulse response (FIR) models there is no way to guarantee stability in the presence of sparsity. Thus we need to utilize a more suitable basis for model expansion.

Discrete time orthonormal Kautz and Laguerre basis functions have gained popularity in the area of system identification. Under mild regularity conditions any transfer function can be represented by an expansion in terms of causal La-

guerre or Kautz polynomials [170, 266]. The z transform of these functions have a recursive structure and the stability of the estimated system can be guaranteed by pre-determining the locations of the poles of these functions. Compact modeling can be achieved by placing the poles close to the dominant poles of the actual system. The benefits of Kautz and Laguerre filters over the delay operator in the context of system identification are given in [5, 102, 170, 266].

Since Kautz and Laguerre basis functions provide a non parametric means of compact modeling of linear dynamic systems with guaranteed stability they have been widely used in system identification [264, 171, 101, 177, 265, 169]. However most of the work has been concentrated on systems with white noise. [121, 221, 196] have developed system identification algorithms with Laguerre expansions that can handle coloured noise but the authors have not introduced sparsity to the estimated system.

Although using Kautz and Laguerre filters for model expansion by itself reduces the order of the approximated model, further reduction in the number of parameters have been accomplished by the incorporation of sparse regression techniques. It has been shown in some applications such as fMRI [29], that sparsity is a natural occurrence when Laguerre basis functions are employed. [22] has used a thresholding method and [189, 190] have used LASSO to reduce the number of parameters used in the Laguerre expansion model. However none of these system identification algorithms can support coloured noise. [214] deals with coloured noise and proposes a l_1 penalized criterion but no algorithm has been presented.

This chapter presents a cyclic descent based algorithm for estimating both a transfer function (modeled as a Laguerre expansion) and a coloured (autoregressive) noise model. Both l_1 and l_0 penalized procedures are discussed to introduce sparsity to the estimated model. The L0LS-CD algorithm developed in section 2.5 is used for the l_0 penalized version and LASSO is used for the l_1 version. Our method can be easily extended to support Kautz filters.

The remainder of the chapter is organized as follows. Section 4.1 gives an overview of the Laguerre filters and introduces the model that will be considered in this chapter. Section 4.2 develops the criterion used for system identification and section 4.3 gives a cyclic descent algorithm to optimize this criterion. Section

4.4 discusses tuning parameter selection. Section 4.5 gives simulation results to demonstrate the effectiveness of this approach and to compare the performance of the two penalties that are considered. Section 4.6 concludes the chapter.

4.1 The Laguerre Models With Coloured Noise

Laguerre basis is a series of filters each comprising of a first order low-pass filter and first order all-pass filters. A k^{th} order discrete time Laguerre basis function is expressed as,

$$\phi_k(q,\gamma) = \frac{\sqrt{1-\gamma^2}}{1-\gamma q^{-1}}\left[\frac{q^{-1}-\gamma}{1-\gamma q^{-1}}\right]^{k-1}, \tag{4.1}$$

where q is the shift operator. γ is a real value and is the decay factor of the Laguerre filter. It must satisfy $|\gamma| < 1$ to ensure stability. It is important to set γ centering the dominant time constants or resonant modes of the system to be identified.

When identifying a system we must first expand it using a proper basis to obtain a parametric representation of its transfer function. Since the Laguerre basis functions form a complete orthonormal set [230] we can expand the transfer function as follows,

$$G(q) = \sum_{1}^{\infty} \beta_k \phi_k(q,\gamma), \tag{4.2}$$

where $G(q)$ is the transfer function to be estimated and β_k are the system parameters. However with this prediction model an infinite number of parameters have to be estimated. To make the task more feasible the representation is truncated to obtain an approximate representation.

$$G(q) \simeq \sum_{1}^{m} \beta_k \phi_k(q,\gamma), \tag{4.3}$$

It has been shown in [15, 211] that with proper selection of γ a truncated Laguerre representation can successfully approximate a system. Note the crucial feature

that aside from the single tuning parameter γ the model is linear in the parameters $\beta_k, k = 1, \cdots, m$.

Consider a system observed in coloured noise:

$$y_t = s_t + n_t, t = 1, \cdots, T, \tag{4.4}$$

where y_t is the measured output, s_t is the system output and n_t is the coloured noise. Using the Laguerre basis expansion the system output can be rewritten as,

$$s_t = \sum_{1}^{m} \beta_k s_{k,t} = x_t^T \beta, \tag{4.5}$$

where $x_t^T = [s_{1,t}, \cdots s_{m,t}]$ and $s_{k,t}, k = 1, \cdots m$ are obtained by filtering the input u_t with Laguerre filters of order $k, k = 1, \cdots m$. The noise model is assumed to be an autoregressive model of order p,

$$n_t = \sum_{l=1}^{p} n_{t-l}\alpha_l + \varepsilon_t = z_t^T \alpha + \varepsilon_t, \tag{4.6}$$

where $z_t^T = [n_{t-1}, \cdots, n_{t-p}]$ and ε_t is white noise. The model (4.3) can be rewritten in a matrix format as follows,

$$y = X\beta + n, \tag{4.7}$$

where $y = [y_1, \cdots, y_T]^T$, $X = [x_1, \cdots, x_T]^T$ and $n = [n_1, \cdots, n_T]^T$. X matrix can also be written as $X = [S_{(1)}, \cdots, S_{(m)}]$, $S_{(k)} = [s_{k,1}, \cdots, s_{k,T}]^T = \phi_k(q, \gamma) * u$ where $u = [u_1, \cdots, u_T]^T$ is the input signal.

The system parameters β and noise model parameters α are unknown and it is our objective to estimate them.

4.2 Sparsity Criterion

Had the system been measured in white noise ($n_t = \varepsilon_t$) then the system parameters could have been estimated by optimizing a penalized least squares criterion

such as (1.4c);

$$J = ||y - X\beta||^2 + hf(\beta), \tag{4.8}$$

where $f(\beta)$ is one of the sparsity measures given in section 1.4. To handle systems with coloured noise, criterion (4.8) has to be modified by replacing the first term with a weighted least squares term.

$$J(\alpha, \beta) = (y - X\beta)^T \Gamma_\alpha^{-1} (y - X\beta) + hf(\beta), \tag{4.9}$$

where Γ_α is the covariance matrix of the noise vector n. To gain a better insight about the criterion (4.9) consider column wise Discrete Fourier Transforms (DFT), $\tilde{y}$, $\tilde{X}$, $\tilde{n}$ of y, X and n respectively. Then (4.7) can be written as $\tilde{y} = \tilde{X}\beta + \tilde{n}$ and the criterion (4.9) can be rewritten approximately as follows,

$$J(\alpha, \beta) = \sum_{k=0}^{T-1} \frac{|\tilde{y}_k - \sum_1^m \tilde{s}_{i,k}\beta_i|^2}{F_k} + hf(\beta), \tag{4.10}$$

where $F_k = F(\omega_k)$, $\omega_k = 2\pi k/T$ is the autoregressive spectrum and $\tilde{S}_{(i)} = [\tilde{s}_{i,1}, \cdots, \tilde{s}_{i,T}]^T$ is the DFT of $S_{(i)}$. The approximation gets better as T gets larger. The spectrum of an autoregressive process of order p is given by,

$$F_k = \frac{\sigma^2}{\left|1 - \sum_{l=1}^p \alpha_l e^{\frac{-j2\pi lk}{T}}\right|^2} = \frac{\sigma^2}{|A_k|^2}. \tag{4.11}$$

Then (4.10) can be rewritten as,

$$J(\alpha, \beta) = \sum_{k=0}^{T-1} \frac{|\tilde{y}_k - \sum_1^m \tilde{s}_{i,k}\beta_i|^2 |A_k|^2}{\sigma^2} + hf(\beta). \tag{4.12}$$

Also,

$$\left|\tilde{y}_k - \sum_1^m \tilde{s}_{i,k}\beta_i\right|^2 |A_k|^2 = \left|A_k\left(\tilde{y}_k - \sum_1^m \tilde{s}_{i,k}\beta_i\right)\right|^2$$
$$= \left|\left(1 - \sum_1^p \alpha_l e^{-jl\omega_k}\right)\left(\tilde{y}_k - \sum_1^m \tilde{s}_{i,k}\beta_i\right)\right|^2$$

Thus the criterion (4.12) can now be rewritten as,

$$J(\alpha,\beta) = L(\alpha,\beta) + hf(\beta), \tag{4.13}$$

where $L(\alpha,\beta) = \sum_t^T\Big[(y_t - \sum_1^p \alpha_l y_{t-l}) - (x_t - \sum_1^p \alpha_l x_{t-l})^T\beta\Big]^2$. Note that $L(\alpha,\beta)$ can be written in two equivalent ways,

$$L(\alpha,\beta) = \sum_{t=1}^T \left[y_{t,f} - x_{t,f}^T\beta\right]^2, \tag{4.14}$$

where $y_{t,f} = y_t - \sum_1^p \alpha_l y_{t-l}$ and $x_{t,f} = x_t - \sum_1^p \alpha_l x_{t-l}$. Also as,

$$L(\alpha,\beta) = \sum_{t=1}^T \left(\hat{n}_t - \sum_1^p \alpha_l \hat{n}_{t-l}\right)^2, \tag{4.15}$$

where $\hat{n}_t = y_t - x_t^T\beta$. These two versions are useful in the cyclic descent algorithm given in section 4.3. In this chapter both $f(\beta) = l_0$ norm (1.2a) and $f(\beta) = l_1$ norm (1.2d) will be considered,

$$J^{(0)}(\alpha,\beta) = L(\alpha,\beta) + h\sum_1^m I(\beta_i \neq 0), \tag{4.16}$$

$$J^{(1)}(\alpha,\beta) = L(\alpha,\beta) + h\sum_1^m |\beta_i|. \tag{4.17}$$

Given y, u and γ the first step is to filter u with Laguerre filters of order $k, k = 1, \cdots, m$ to get $S_{(k)}$. However for this we need to know the initial conditions of the system $(u_0, \cdots u_{-(m-1)})$. Similarly at (4.15) we need to know the initial values of system noise $(n_0, \cdots n_{-(p-1)})$. Since these are generally unknown, after

calculating $S_{(k)}$ or $\hat{n}_t$, disregard the first $d = \max(m, p)$ data points and use only the data points at $t = d, \cdots, T$ in the calculations.

4.3 SCSI Algorithm

Cyclic descent based SCSI (sparse coloured system identification) algorithm is presented here to optimize criterion $J^{(0)}(\alpha, \beta)$ and $J^{(1)}(\alpha, \beta)$. Two version of this algorithm are discussed; SCSI-L0 optimizes $J^{(0)}(\alpha, \beta)$ and SCSI-L1 optimizes $J^{(1)}(\alpha, \beta)$. These criterion have to be optimized with respect to two vector variables α and β. Thus the cyclic descent algorithm has two steps; α-step and β-step. There is a closed form solution for the α-step but the solution for the β-step requiters an iterative procedure.

4.3.1 SCSI-L0

α-step: Given β_{k-1} get α_k

With β fixed we can calculate $\hat{n}_t$. Then $L(\alpha, \beta)$ is given by (4.15) and $\arg.\min_\alpha J^{(0)}(\alpha, \beta_{k-1}) = \arg.\min_\alpha L(\alpha, \beta_{k-1})$. This problem is a linear regression and α_k is the least squares solution of (4.15).

β-step: Given α_k get β_{k+1}

With α fixed we can filter y_t and x_t, $t = 1, \cdots, T$ to get $y_{f,t}$ and $x_{f,t}$. Then $L(\alpha, \beta)$ is given by (4.14). $J^{(0)}(\alpha_k, \beta)$ becomes an l_0 penalized least squares criterion,

$$J^{(0)}(\alpha_k, \beta) = ||y^f - X^f\beta||^2 + h\sum_1^m I(\beta_i \neq 0),$$

where $y^f = [y_{1,f}, \cdots, y_{T,f}]^T$ and $X^f = [x_{1,f}, \cdots, x_{T,f}]^T$. This can be solved by the L0LS-CD algorithm given in section 2.5.

4.3.2 SCSI-L1

α-step: Given β_{k-1} get α_k

This step is the same as the α-step of SCSI-L0. Given β we can calculate $\hat{n}_t$. Then $L(\alpha, \beta)$ is given by (4.15). $\arg.\min_\alpha J^{(1)}(\alpha, \beta_{k-1}) = \arg.\min_\alpha L(\alpha, \beta_{k-1})$. Thus α_k is the least squares solution of (4.15).

β-step: Given α_k get β_{k+1}

With α fixed calculate $y_{f,t}$ and $x_{f,t}$. Then $L(\alpha, \beta)$ is given by (4.14). $J^{(1)}(\alpha_k, \beta)$ becomes an l_1 penalized least squares criterion,

$$J^{(1)}(\alpha_k, \beta) = ||y^f - X^f\beta||^2 + h\sum_1^m |\beta_i|,$$

This can be solved by the LASSO [6, 236] algorithm.

To initialize the SCSI algorithm we need β_0. For this, assume that the noise is white (set $\alpha = 0$) and find the least squares solution of (4.14). Since the expected value of the coloured noise is zero ($E[n] = 0$) this is an unbiased estimate of β.

Once β_0 is found, initialize the iteration count $k = 1$ and sequentially repeat α-step and β-step until the termination criterion is met. Iteration count has to be incremented by one after each step.

$J^{(0)}(\alpha_k, \beta)$ has many local minimum unlike $J^{(1)}(\alpha_k, \beta)$. So unlike L0LS-CD, the estimate of LASSO does not depend on the initialization. Thus proper initialization is essential to guarantee minimization when optimizing $J^{(0)}(\alpha_k, \beta)$. Initialize L0LS-CD with β_{k-1} at the k^{th} iteration to ensure that the criterion reduces its value at each β-step.

To terminate this algorithm, monitor the value of the criterion at each iteration and terminate when $J_k - J_{k-1}$ <tolerance.

4.4 Tuning Parameter Selection

Many parameters need to be known in order to use the SCSI algorithm. Generally system model order (m), noise model order (p), appropriate value for the decay parameter (γ) and penalty parameter (h) are unknown.

Since l_0 or l_1 penalized least squares criterion is used to estimate the system parameters (β), m can be set to its upper bound. L0LS-CD and LASSO promotes

sparsity, thus it will zero out any unnecessary parameters of the β vector. The actual order of the system can be identified by ignoring the zeros at the end of the final estimate of β. p is also set to its upper bound.

To find out the best values of h and γ, run the SCSI algorithm for a range of $[h, \gamma]$ combinations. The best combination is found by the values which minimizes the BIC criterion. Since the variance of noise is unknown in this application, variant of (2.44) is used as the BIC criterion,

$$\text{BIC} = \ln\left(\frac{L(\hat{\alpha}_{h,\gamma}, \hat{\beta}_{h,\gamma})}{T-d}\right) + \frac{r}{T-d}\ln(T-d), \tag{4.18}$$

where $\hat{\alpha}_{h,\gamma}$ and $\hat{\beta}_{h,\gamma}$ are the output of the SCSI algorithm and r is the number of non zero coefficients in $\hat{\beta}_{h,\gamma}$.

The dominant pole of a system can be identified with the knowledge of the impulse response of the system. If a rough estimate of the suitable value of γ is known a priori, the search can be restricted to its vicinity rather than searching the whole range $(-1 < \gamma < 1)$ for a suitable value. Range of h is $[0, \tilde{h}]$ where $\tilde{h}$ is the minimum value of h that sets all β coefficients to zero.

4.5 SCSI Simulation

This section presents simulation results to demonstrate the performance of the SCSI algorithm. To compare the performance of the l_0 penalty and the l_1 penalty, both SCSI-L0 and SCSI-L1 algorithms were applied on the same set of data. Performance measures MSE_β (2.40), TPR (2.41) and FPR (2.42) defined in section 2.11.1 is used in this application.

Consider a linear system represented by a Laguerre filter of order 7 with the following parameters $[1, 0.5, 0, 0.7, 0.1, 0, 0.6]$ and set the decay factor γ to 0.5. The system was excited with white Gaussian input with zero mean and unit variance ($u \sim N(0, 1)$) and 250 data points ($T = 250$) were considered. Due to the recursive structure of the Laguerre filters the system model can be efficiently implemented as shown in figure 4.1 where $L_0(q, \gamma) = \sqrt{1-\gamma^2}/1 - \gamma q^{-1}$ and $H(q, \gamma) = q^{-1} - \gamma/1 - \gamma q^{-1}$.

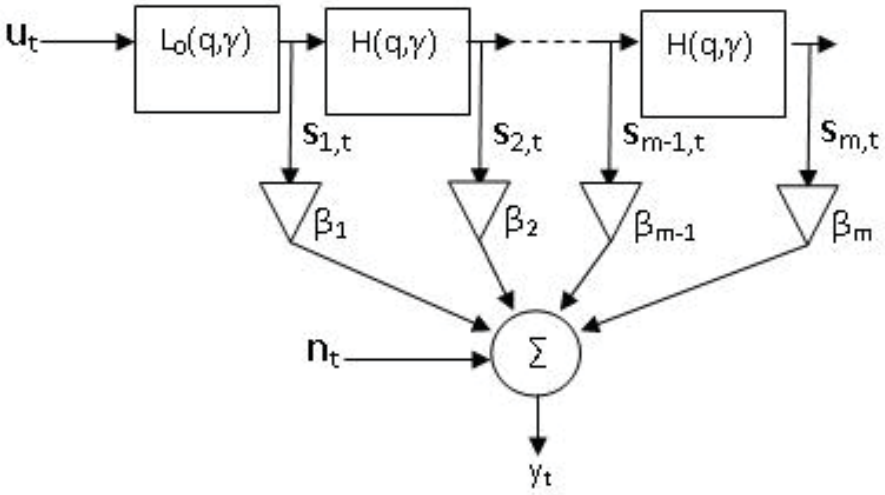

Figure 4.1: Block diagram of an efficient implementation of the system model by exploring the recursive nature of Laguerre filters.

The noise model is assumed to be an autoregressive model of order 3 with parameters $[1.2, -0.4625, 0.05625]$ such that its characteristic polynomial has roots $0.5, 0.45, 0.25$, which are within the unit disk to ensure stability. The noise model was excited with white Gaussian input with zero mean and σ^2 variance such that the SNR is 10. SNR is given by,

$$\text{SNR} = \frac{||X\beta||^2}{||n||^2}. \tag{4.19}$$

Since the order of the system is unknown it was set to double the size of that of the actual system ($m = 14$) and since the noise model order is assumed to be known, p was set to 3. Since a rough idea about the location of the dominant pole of the system is to be known a priori, γ was varied in the range of $[0.3, 0.7]$ in steps of 0.025. As described in section 4.4 the best value for h and γ are those that produces estimates which minimize the BIC criterion (4.18).

The estimates of SCSI-L0 and SCSI-L1 are compared with the actual values of the system and noise model parameters in table 4.1. We can clearly see that SCSI-L1 estimates have many incorrectly identified non-zeros (high FPR). This observation with regards to the l_1 penalty has already been made in sections 2.11.5, 2.12.2 and in many other occasions [203, 189, 190]. In contrast SCSI-L0

algorithm sets more coefficients to zero and the locations of the non-zeros align with that of the original system (FP= 0). The estimated values of β, α and γ by SCSI-L0 are very similar to the actual values of the system. Further more the actual order of the system can be recovered by discarding the 7 zeros at the end of $\hat{\beta}$. However the actual order of the system cannot be obtained by SCSI-L1 estimates since even $\hat{\beta}_{13}$ is non zero.

Table 4.1: Comparison of the estimates of SCSI-L0 and SCSI-L1 with the actual system and noise parameters

	Actual	SCSI-L0 Estimate	SCSI-L1 Estimate
β_1	1	1.0089	0.9885
β_2	0.5	0.5249	0.4282
β_3	0	0	-0.0286
β_4	0.7	0.6741	0.6321
β_5	0.1	0.1184	0
β_6	0	0	0.0471
β_7	0.6	0.5687	0.5074
β_8	0	0	-0.1318
β_9	0	0	0
β_{10}	0	0	-0.0121
β_{11}	0	0	0
β_{12}	0	0	0
β_{13}	0	0	0.0205
β_{14}	0	0	0
α_1	1.2	1.1758	1.1731
α_2	-0.4625	-0.4735	-0.4625
α_3	0.05625	0.0837	0.0763
γ	0.5	0.5	0.525

Figure 4.2 shows the variation of the coefficients of the $\hat{\beta}$ vector as a function of $\hat{\gamma}$ in SCSI-L0. We can clearly see that the number of non zero coefficients reduce dramatically at the vicinity of $\gamma = 0.5$. This further demonstrates that proper selection of the decay parameter γ assists in the compact representation of the system when using Laguerre functions for model expansion.

From these results we can clearly see that the SCSI algorithm can successfully

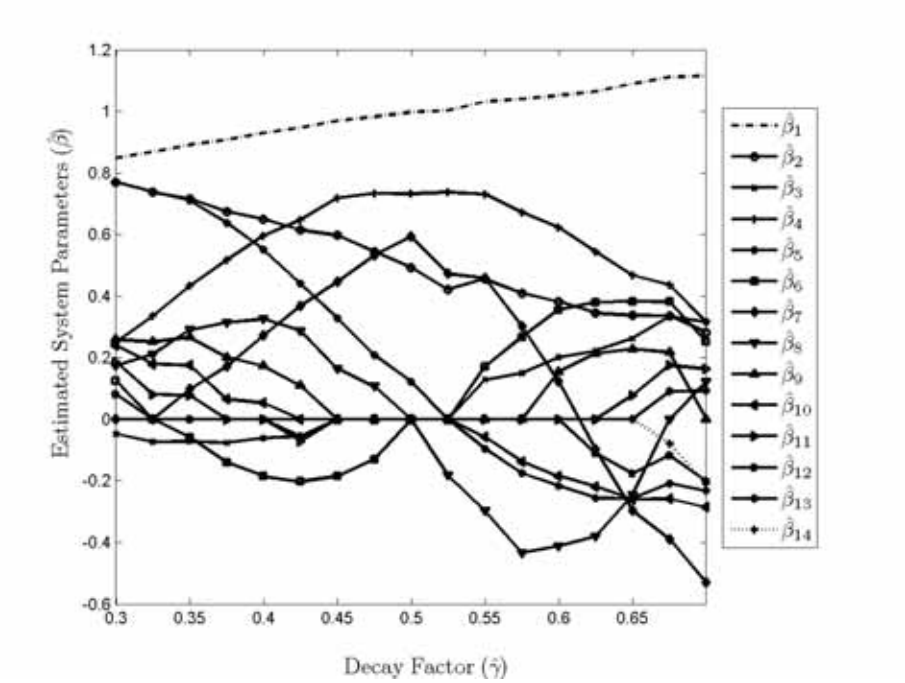

Figure 4.2: Variation of the estimated system parameters ($\hat{\beta}$) by SCSI-L0 as a function of the decay factor ($\hat{\gamma}$).

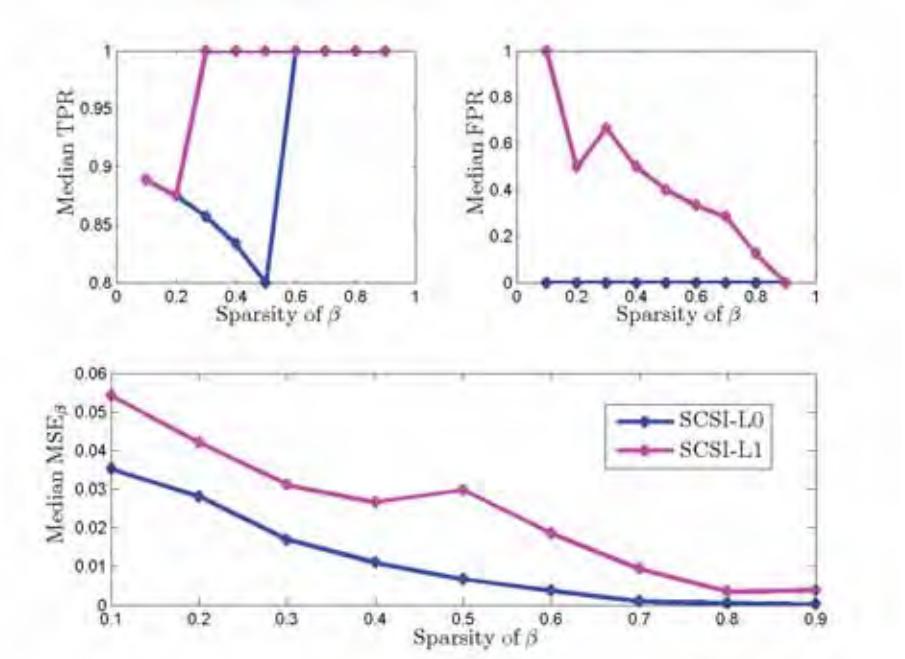

Figure 4.3: Performance comparison of SCSI-L0 and SCSI-L1 by the variation of performance measures MSE_β, TPR and FPR as a function of sparsity of $\beta^\star$.

identify the system model and the noise model when the data is corrupted by coloured noise. Furthermore these simulation results are evidence that the l_0 penalty promotes more sparsity and is more suitable than the l_1 penalty.

The performance of SCSI-L0 and SCSI-L1 was further analyzed at different levels of sparsity of $\beta^\star$. Sparsity was varied from 0.1 to 0.9 in steps of 0.1 and 150 $\beta^\star$ vectors were generated per sparsity level by placing non zeros entries drawn from a uniform distribution on the open interval $(0, 1)$ at random locations of a zero vector of length 10. The α vector, SNR and γ were kept the same as the previous example and the input signal and the noise was also generated as stated above. m was set to 10 to maintain the expected sparsity of the estimates. The median of the performance indicators of the estimates of SCSI-L0 and SCSI-L1 are given in figure 4.3.

Although the TPR of SCSI-L1 is higher than that of SCSI-L0 it has also got a very high FPR. Furthermore the MSE_β of SCSI-L0 is lower than that of SCSI-L1. So the estimates of SCSI-L1 are non sparse and have a higher estimation error and therefore SCSI-L0 is more desirable than SCSI-L1.

4.6 Conclusion

This chapter developed a cyclic descent based system identification algorithm that recovers sparse transfer functions based on Laguerre basis functions with system output measured in coloured noise. Unlike most other sparse approaches to transfer function estimation the Laguerre approach guarantees stability. Two competing criterion were developed based on the l_0 penalty and the l_1 penalty. Simulation results show that the algorithm successfully recovers the system and the noise models and that the performance of the criterion with l_0 penalty is superior to that with the l_1 penalty.

Chapter 5

Application: Sparse Network Topology Identification

The problem of identifying the topology of a network of dynamic systems from a time series data arises in applications in a vast array of disciplines such as, economics [165], thermal dynamics [178], epidemiology [160], ecology [254], geology [12], sociology [268] and in biological examples such as metabolism [261], genetic networks [75] and protein interaction networks [248]. Due to the growing demand many methods of network topology identification have been developed [23, 99, 11, 197, 151, 152].

Causal relationships between time series was first studied by Granger [92] and has been widely used in economics since. The basic concept of Granger causality is that if one time series (y) is caused by another time series (x) then the knowledge of the past values of x improves the prediction of y compared to only using past values of y. Granger causality of multivariate time series was first explored within the framework of autoregressive models in electrophysiological signal analysis [82, 83]. This was further developed with a graphical framework in [58].

In a network of N dynamic systems, suppose that the output of each system can be observed and that each corresponds to a time series. It is assumed that the time series is a result of an underlying, unknown directed graph topology. The output of each dynamic system is causally effected by the incoming data of the

directly connected nodes and system noise. If it is assumed that the network is fully connected (every node causally influences every other node) the result would be a very flexible model but it would also be over parameterized and estimation problem will be ill-conditioned. Normally a network of dynamic system is sparsely connected (a node is effected by only a few other nodes in the network), thus an additional criterion of sparsity is incorporated to the estimation problem.

[256, 10, 181, 280] have utilized sparse regression techniques based on the l_1 norm to identify a sparse network. However penalizing individual parameters of the network model will result in many links each with few coefficients being selected. To successfully recover a sparse network, parameters corresponding to each link should be grouped and algorithms that focus on group sparsity should be employed.

The group LASSO [278] algorithm was used to recover a sparse network in [23, 99] and [11] have reformulated the problem with a re-weighted iterative procedure. Group sparse versions of greedy algorithms were used in [197, 151, 152] to recover sparse networks. Clustered orthogonal matching pursuit was used in [197] and cycling orthogonal least squares was used in [151, 152]. However all these methods are applied using an autoregressive model framework.

Very high order autoregressive models are required to approximate systems with rapid sampling. This issue is overcome by the usage of causal Laguerre polynomial basis expansion to model transfer functions between two time series. An overview of causal Laguerre basis functions is given in chapter 4.

In this chapter a network identification method with Laguerre basis functions is developed. The sparse network is recovered with the use of gL0LS-CD algorithm which was developed in section 2.9. gL0LS-CD is compared with group LASSO in the simulations.

The remainder of this chapter is organized as follows. Section 5.1 introduces the network model and section 5.2 develops the sparsity criterion for sparse network recovery. Simulation results to demonstrate the effectiveness of this method are given in section 5.3 and conclusions are in section 5.4.

5.1 Network Model

A network of dynamic systems can be represented by a directed graph as shown in figure 5.1. Each dynamic system is represented by a node and each causal dependance is represented by a link with an arrow. The output (time series) of each dynamic system can be observed. Each time series (x^i) can be modeled as follows,

$$x_t^i = G_{i,i}(q)x_{t-1}^i + \sum_{j \in A_i} G_{i,j}(q)x_{t-d_{i,j}}^j + e_t^i, \tag{5.1}$$

where x_t^i and e_t^i are the output and system noise of node i at time t respectively, $G_{i,j}(q)$ and $d_{i,j}$ are the transfer function and transmission delay between node i and j respectively, q is the forward shift operator and A_i is the set of indices of the parent nodes of node i. A graphical representation of the model of a single node is given in figure 5.2.

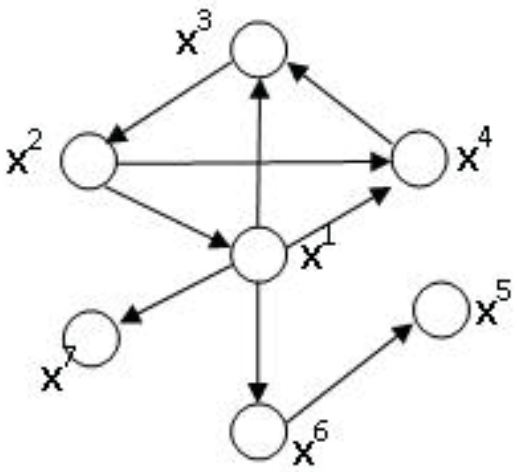

Figure 5.1: Directed graph of a network of 7 dynamic systems

Since the topology of the network is unknown we do not know A_i of (5.1). Thus we assume that the network is fully connected and use sparse regression

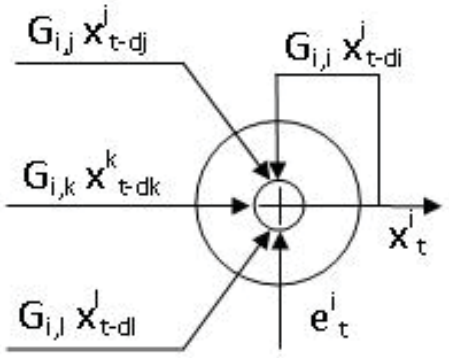

Figure 5.2: Model of a single node within a network of dynamic systems

techniques to recover A_i. Thus the general model of a node is given by,

$$x_t^i = \sum_{j=1}^{N} G_{i,j}(q) x_{t-d_{i,j}}^j + e_t^i, \tag{5.2}$$

where N is the total number of nodes in the network. Transfer functions $G_{i,j}$ has to be expanded using a proper basis. So far in the literature the delay operator is used as the basis. Then equation (5.2) can be rewritten as,

$$x_t^i = \sum_{j=1}^{N} \sum_{k=1}^{m} \beta_{i,j,k} q^{-k} x_{t-d_{i,j}}^j + e_t^i. \tag{5.3}$$

Here m is set to its upper bound such that the highest order transfer function of the network can be accommodated. An introduction to model expansion with Laguerre basis functions is given in section 4.1. Thus the equation (5.2) can be rewritten with the Laguerre basis expansion as follows,

$$x_t^i = \sum_{j=1}^{N} \sum_{k=1}^{m} \beta_{i,j,k} \phi_k(q, \gamma_{i,j}) x_{t-d_{i,j}}^j + e_t^i. \tag{5.4}$$

Equations (5.3) and (5.4) can be written in matrix regression form as follows,

$$y^i = X^i \beta^i + e^i, \tag{5.5a}$$

where

$$y^i = [x_t^i, \cdots, x_{t-p}^i]^T, \tag{5.5b}$$
$$\beta^i = [\beta_{i,1,1}, \cdots, \beta_{i,1,m}, \beta_{i,2,1}, \cdots, \beta_{i,N,m}]^T, \tag{5.5c}$$
$$e^i = [e_t^i, \cdots, e_{t-p}^i]^T, \tag{5.5d}$$

also X^i contains filtered signals as follows,

$$X^i = [H_{i,1}, H_{i,2}, \cdots, H_{i,N}], \tag{5.5e}$$
$$H_{i,j} = \begin{bmatrix} \varphi_{i,j,1,t} & \varphi_{i,j,2,t} & \cdots & \varphi_{i,j,m,t} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{i,j,1,t-p} & \varphi_{i,j,2,t-p} & \cdots & \varphi_{i,j,m,t-p} \end{bmatrix}. \tag{5.5f}$$

If the delay operator was used then,

$$\varphi_{i,j,k,t} = q^{-k} x_{t-d_j}^j = x_{t-d_j-k}^j, \tag{5.5g}$$

else, if the Laguerre basis functions were used then,

$$\varphi_{i,j,k,t} = \frac{\sqrt{1-\gamma_{i,j}^2}}{1-\gamma_{i,j}q^{-1}} \left[\frac{q^{-1}-\gamma_{i,j}}{1-\gamma_{i,j}q^{-1}}\right]^{k-1} x_{t-d_j}^j. \tag{5.5h}$$

5.2 Sparsity Criterion

In (5.5c) the coefficients corresponding to each link can be grouped[1] as $\beta^i = [\beta_1^{i,T}, \cdots \beta_N^{i,T}]^T$ where $\beta_j^i = [\beta_{i,j,1}, \cdots, \beta_{i,j,m}]^T$,. Each column-wise partition $H_{i,j}$ of X^i in (5.5e) corresponds to the β_j^i groups. Here all the groups are of the same size but it need not be so. If the N partitions of β^i is of size m_j then the length of β^i is $n = \sum_{j=1}^N m_j$, and the partitions of X^i, $H_{i,j}$ will have the corresponding number of m_j columns. Similar to section 2.8, the group l_0 penalized least squares

[1] Should not be confused with the notation used in chapter 2 for grouped variables.

criterion for (5.5a) can be given as follows,

$$J^{(0)}(\beta^i) = ||y^i - X^i\beta^i||^2 + h\sum_{j=1}^{N} I(||\beta_j^i|| \neq 0), \tag{5.6}$$

gL0LS-CD algorithm developed in section 2.9 can be used to optimize $J^{(0)}(\beta^i)$. Similarly the group l_1 penalized least squares criterion for (5.5a) is as follows,

$$J^{(1)}(\beta^i) = ||y^i - X^i\beta^i||^2 + h\sum_{j=1}^{N} ||\beta_j^i||. \tag{5.7}$$

Here the penalty is a mixture of l_1 and l_2 norms and is the sum of the lengths of the partitions of β^i. The group LASSO algorithm given in [278] can be used to optimize $J^{(1)}(\beta^i)$. An overview of group LASSO will be given in section 6.2.1.

The Criteria $J^{(0)}(\beta^i)$ and $J^{(1)}(\beta^i)$ promote group sparsity but do not encourage sparsity within the groups. Once the topology of the network is identified the transfer function can be further refined by the following criterion,

$$\bar{J}^{(0)}(\bar{\beta}^i) = ||y^i - \bar{X}^i\bar{\beta}^i||^2 + h\sum_{j=1}^{\bar{n}} \|\bar{\beta}_j^i\|_0, \tag{5.8}$$

or by,

$$\bar{J}^{(1)}(\bar{\beta}^i) = ||y^i - \bar{X}^i\bar{\beta}^i||^2 + h\sum_{j=1}^{\bar{n}} \|\bar{\beta}_j^i\|_1, \tag{5.9}$$

where $\bar{\beta}^i$ is composed of $\beta_j^i, j \in A_i$, $\bar{X}^i$ is composed of $H_{i,j}, j \in A_i$, $\bar{n} = |A_i|$ and $|A_i|$ is the cardinality of the set. $\bar{J}^{(0)}$ can be optimized by the L0LS-CD algorithm given in section 2.5 and similarly $\bar{J}^{(1)}$ can be optimized by LASSO [6, 236].

5.3 Network Topology Simulation

This section simulate a network of interconnected dynamic systems to generate time series data and then estimate the network topology using the identification method developed in this chapter. To compare the performance of the l_0 and l_1

penalties, both $J^{(0)}(\beta^i)$ and $J^{(1)}(\beta^i)$ criteria were optimized on the same set of data. Two networks studied in previous literature [23] are considered here and they are given in figure 5.3. The output of each node is driven by its own past values as well as the output of its parent nodes.

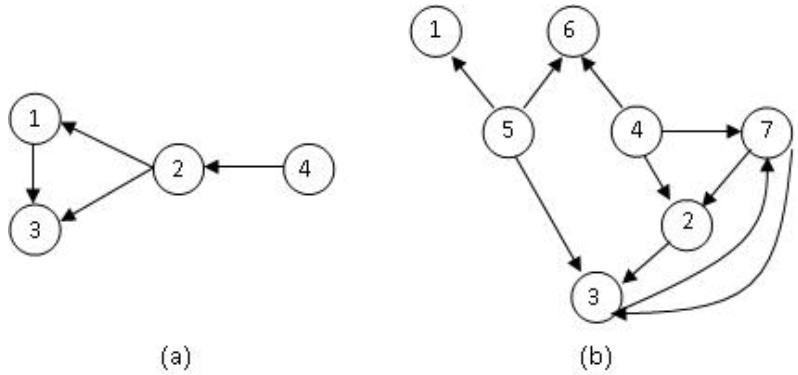

Figure 5.3: Network topologies from existing literature.

To implement the network each transfer function was represented using an autoregressive model. Order of all the models were set to 5 and the coefficients were taken from a Gaussian normal distribution with zero mean and 0.04 variance. Only the stable realizations were selected.

The networks were excited with Gaussian random noise with zero mean and 0.01 variance. 30 realizations of such networks were generated with $p = 300$ time samples recorded at each node. The time series were then filtered to generate the $X^i, i = 1, \cdots, N$ matrices as given in (5.5e)-(5.5h). Transmission delays $d_{i,j}$ and decay factors of the Laguerre filters $\gamma_{i,j}$ have to be known for (5.5h). $d_{i,j}$ was set to 1 and all the decay factors were set to a fixed value for the whole network $\hat{\gamma}$. Any mismatch in the decay factor will simply result in few additional terms being included in the estimated system. Since the appropriate value for $\hat{\gamma}$ is unknown a range of values $-1 < \hat{\gamma} < 1$ were considered and the best value for $\hat{\gamma}$ was selected using the BIC criterion. Since the variance of noise is unknown, a variant of (2.44) is used as the BIC criterion,

$$\text{BIC} = \ln\left(\sum_{i=1}^{N} \frac{\|y^i - X^i \beta^i_{\hat{\gamma},h}\|^2}{p}\right) + \frac{r}{pN} \ln(pN),$$

where r is the total number of non zero coefficients in all the $\beta^i_{\hat{\gamma},h}$ vectors. For

each value of $\hat{\gamma}$, $X^i, i = 1, \cdots, N$ matrix was generated and gL0L-CD and group LASSO were applied to individual nodes. A range of values were considered for the penalty parameter $0 < h < \hat{h}$, where $\hat{h}$ is the minimum value of h that sets all the estimates to zero ($\beta^i = 0$). The best value of h for each node was selected with BIC. Variation of BIC of a single node with respect to h and $\hat{\gamma}$ is given in figure 5.4. As shown in figure 5.4, BIC is minimized in the vicinity of $\gamma = 0$ which is expected as the original system was based on a multivariate autoregressive model.

The probability of each link of networks 'a' and 'b' being identified by gL0LS-CD and group LASSO are given in tables 5.1 and 5.2 respectively.

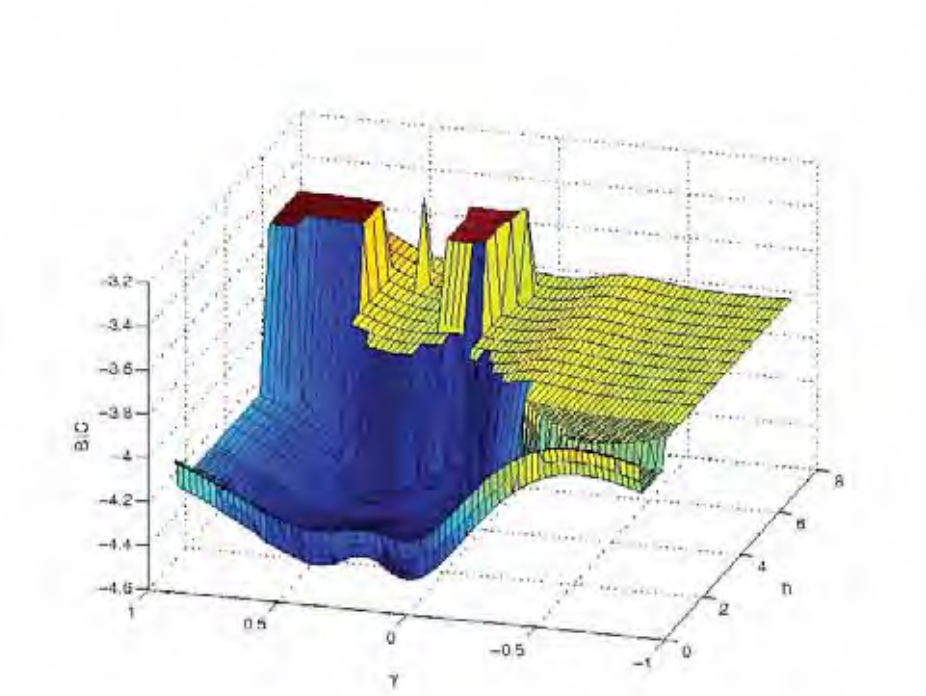

Figure 5.4: Variation of BIC criterion at a single node as a function of the decay factor ($\hat{\gamma}$) and the penalty parameter (h).

The 8 links that exsist in network 'a' are given by the first 8 rows of table 5.1 and both gL0LS-CD and group LASSO successfully identify these links with high probability. However from the last 3 rows we can see that group LASSO selects more links than what actually exist in the network. gL0LS-CD does not show this property (FP = 0).

Table 5.1: Probability of each link of network 'a' being identified by gL0LS-CD and group LASSO

Link	gL0LS-CD (%)	group LASSO(%)
$1 \rightarrow 1$	90	83.33
$2 \rightarrow 1$	90	90
$2 \rightarrow 2$	73.33	73.33
$4 \rightarrow 2$	83.33	76.67
$1 \rightarrow 3$	93.33	93.33
$2 \rightarrow 3$	86.67	90
$3 \rightarrow 3$	70	86.67
$4 \rightarrow 4$	80	70
$3 \rightarrow 1$	0	6.67
$4 \rightarrow 1$	0	10
$4 \rightarrow 3$	0	53.33

The links that exist in the original network 'b' are given in table 5.2. Similar to the earlier example both algorithms identify these links with high probability. However similar to the earlier observation, group LASSO selects links that does not exist in the original system such as '$3 \rightarrow 2$', '$6 \rightarrow 2$', '$1 \rightarrow 3$', '$4 \rightarrow 3$', '$6 \rightarrow 3$', '$2 \rightarrow 6$', '$7 \rightarrow 6$', '$1 \rightarrow 7$', '$2 \rightarrow 7$', '$5 \rightarrow 7$', '$6 \rightarrow 7$' with probabilities ranging from 3.33% to 16.67%. gL0LS-CD does not select any link that does not exist in the original network.

This tendency of the l_1 penalty to have high false positives in its estimates was previously observed in sections 2.11.5, 2.12.2, 4.5. [23] have derived conditions under which group LASSO consistently estimates the sparse network. From the simulation results we can clearly see that gL0LS-CD is superior to group LASSO as its estimates are more sparse and as it is less likely to select links that does not exist in the original network. In other words gL0LS-CD works even when group LASSO fails.

Table 5.2: Probability of each link of network 'b' being identified by gL0LS-CD and group LASSO

Link	gL0LS-CD (%)	group LASSO(%)
$1 \rightarrow 1$	90	83.33
$5 \rightarrow 1$	73.33	76.67
$2 \rightarrow 2$	80	80
$4 \rightarrow 2$	76.67	70
$7 \rightarrow 2$	93.33	90
$2 \rightarrow 3$	80	90
$3 \rightarrow 3$	93.33	93.33
$5 \rightarrow 3$	73.33	76.67
$7 \rightarrow 3$	96.67	93.33
$4 \rightarrow 4$	76.67	53.33
$5 \rightarrow 5$	83.33	66.67
$4 \rightarrow 6$	90	70
$5 \rightarrow 6$	80	70
$6 \rightarrow 6$	90	66.67
$3 \rightarrow 7$	96.67	90
$4 \rightarrow 7$	90	86.67
$7 \rightarrow 7$	76.67	86.67

5.4 Conclusion

This chapter discusses a method for topology identification of a sparse network of dynamic systems from the time series data collected by measuring the output of each node in the presence of noise. The transfer function were expanded using Laguerre basis functions and gL0LS-CD algorithm was used to estimate the network topology. BIC was used to select the penalty parameter h and the decay parameter γ. It is clear from the simulation results that the method presented in this chapter successfully identifies the topology of the network and gL0LS-CD is superior to group LASSO in recovering sparse estimates of a linear inverse problem with grouped variables.

Chapter 6

Parameter Based Model Selection via SURE

The optimization criterion (1.4a) - (1.4c) used in sparse approximation involves the selection of a tuning/penalty parameter. The value of the penalty parameter determines the sparsity of the estimate. Many competing models can be derived by varying the value of the penalty parameter.

As mentioned in the section 1.8 the performance of the competing models are measured using some discrepancy measure. However discrepancy measures generally cannot be directly calculated. Thus model selection criteria have been developed to select the value of the penalty parameter that corresponds to the best model with respect to some discrepancy measure. Model selection criteria typically select this value by minimizing an estimate of a selected discrepancy measure.

An overview of model selection criteria is given in section 1.8 and SURE stands out as a model selection criterion which does not suffer from the limitations of the other model selection methods. SURE was introduced by [226] and has been used in a wide range of applications as a tuning parameter selection method [208, 215, 217, 218, 250, 219, 138, 149, 61]. All the model selection criteria discussed in section 1.8 focuss on minimizing signal or prediction mean squared error (MSE_μ) (2.39). This chapter develops a model selection criterion for over-determined systems based on parameter mean squared error (MSE_β) (2.40).

As mentioned in section 1.5 LASSO which solves the l_1 penalized least squares criterion is one of the most commonly used algorithms in sparse regression. The importance of selecting a proper penalty parameter when solving this criterion has had little treatment. The SURE criterion for MSE_μ for the special case of orthogonal regressors was developed in [278] and extended in [56]. SURE for MSE_μ for the LASSO was presented in [283] while [223] derives SURE for MSE_μ for the group LASSO. This chapter develops SURE for MSE_β for the group LASSO. This seems to be the first time SURE has been used to deal with MSE_β.

The remainder of this chapter is organized as follows. Section 6.1 develops the general SURE criterion for MSE_β. In section 6.2 this result is applied to obtain SURE for MSE_β for the l_1 penalized least squares problem with grouped variables. Section 6.3 presents simulation results based on a sparse network application, comparing the SURE criterion for MSE_β with other model selection criteria. The application used in this simulation was introduced in chapter 5. The conclusion is in section 6.4.

6.1 SURE for Parameter Mean Squared Error

Consider the linear regression model (1.1) introduced in section 1.1. Denote $\mu_{n\times 1} = X_{n\times p}\beta_{p\times 1}$ as the noise free signal and consider the noise to be Gaussian with zero mean $\varepsilon \sim N(0, \sigma^2 I)$. In the context of interest of this thesis, given y and X, the parameter vector β is estimated using a regularization method which requires the selection of a penalty or tuning parameter (h). For a given value of h denote $\hat{\beta}_h$ and $\hat{\mu}_h = X\hat{\beta}_h$ as the estimated coefficient/parameter vector and estimated signal respectively. Then the signal mean squared error is given by,

$$\text{MSE}_\mu = R_{\mu,h} = E\|\mu - \hat{\mu}_h\|_K^2, \tag{6.1}$$

where K is a given weighting matrix. The parameter mean squared error is given by,

$$\text{MSE}_\beta = R_{\beta,h} = E\|\beta - \hat{\beta}_h\|_\Gamma^2, \tag{6.2}$$

where Γ is also a given weighting matrix. It should be noted that when MSE_μ and MSE_β were used as performance measures for sparse estimation they were normalized, however they need not be so when used for model selection. We have to find the h that minimizes $R_{\beta,h}$. Since β is unknown it is not possible to directly calculate $R_{\beta,h}$ so the idea is to find an unbiased estimator of $R_{\beta,h}$ and minimize that instead.

We now proceed to a derivation of SURE for $R_{\beta,h}$. SURE relies on representing the data in a signal plus noise form e.g. $y = \mu + \varepsilon$. Then SURE produces an unbiased estimator of $R_{\mu,h} = E\|\mu - \hat{\mu}_h\|^2$. So this traditional setup would not work for our purposes.

The problem has to be transformed so that the signal becomes β. To do this, begin with the least squares estimator of an over-determined system,

$$\begin{aligned} z &= (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + \varepsilon), \\ &= \beta + (X^T X)^{-1} X^T \varepsilon, \\ &= \beta + w, \end{aligned} \tag{6.3}$$

where $w \sim N(0, \Omega)$, $\Omega = \sigma^2 (X^T X)^{-1}$. This transformation has achieved our aim but at a cost. We now have pseudo-data $z = \text{signal}(\beta) + \text{noise}(w)$ but the noise has a covariance matrix $\sigma^2 (X^T X)^{-1}$.

We now proceed to derive SURE for this new model. Add and subtract z inside (6.2) to obtain,

$$R_{\beta,h} = E\|z - \hat{\beta}_h - (z - \beta)\|_\Gamma^2 = E\|e_h - w\|_\Gamma^2, \tag{6.4}$$

where $e_h = z - \hat{\beta}_h$. Expanding gives,

$$R_{\beta,h} = E\|e_h\|_\Gamma^2 - 2E(e_h^T \Gamma w) + E\|w\|_\Gamma^2. \tag{6.5}$$

Now $E\|w\|_\Gamma^2$ does not depend on h so it can be dropped. An estimator of the first term is $\|e_h\|_\Gamma^2$. The middle term can be expanded as,

$$E(e_h^T \Gamma w) = E(e_h^T \Gamma (z - \beta)) = \int e_h^T \Gamma (z - \beta) p(z) dz, \tag{6.6}$$

where $p(z)$ is the probability density of z. Since $w \sim N(0, \Omega)$, from (6.3) $z \sim N(\beta, \Omega)$, thus

$$p(z) = \frac{e^{-\frac{1}{2}(z-\beta)^T \Omega^{-1}(z-\beta)}}{(2\pi)^{\frac{n}{2}} |\Omega|^{\frac{1}{2}}}, \tag{6.7}$$

so that $\partial p/\partial z = -\Omega^{-1}(z-\beta)p(z)$. Thus we can write,

$$E(e_h^T \Gamma w) = -\int e_h^T \Gamma \Omega \frac{\partial p}{\partial z} dz. \tag{6.8}$$

Integrating by parts gives

$$E(e_h^T \Gamma w) = -[e_h^T \Gamma \Omega p(z)]_{-\infty}^{\infty} + \int \text{trace}\left(\Gamma \Omega \frac{\partial e_h^T}{\partial z}\right) p(z) dz. \tag{6.9}$$

Since $p(z)$ decays exponentially it is reasonable to assume that $e_h^T p(z)$ vanishes at $\pm\infty$. Thus the first term of (6.9) vanishes and we can obtain,

$$\begin{aligned} E(e_h^T \Gamma w) &= \int \text{trace}\left(\Gamma \Omega \frac{\partial e_h^T}{\partial z}\right) p(z) dz, \\ &= E\left(\text{trace}\left(\Omega \Gamma \frac{\partial e_h}{\partial z^T}\right)\right). \end{aligned} \tag{6.10}$$

So introducing $\hat{R}_{\beta,h} = \|e_h\|_\Gamma^2 - 2\left(\text{trace}\left(\Omega\Gamma \frac{\partial e_h}{\partial z^T}\right)\right)$, we see that $\hat{R}_{\beta,h} + \|w\|_\Gamma^2$ is an unbiased estimator of $R_{\beta,h}$ and we can choose h as the minimizer of $\hat{R}_{\beta,h}$. $\hat{R}_{\beta,h}$ is the SURE for $R_{\beta,h}$. From the definition of e_h we can also write,

$$\hat{R}_{\beta,h} = \|e_h\|_\Gamma^2 + 2\left(\text{trace}\left(\Omega\Gamma \frac{\partial \hat{\beta}_h}{\partial z^T}\right)\right) - 2\text{trace}(\Omega\Gamma) \tag{6.11}$$

The last term does not depend on h and so can be dropped. Still the resulting expression will be referred to as SURE for $R_{\beta,h}$. When $\Gamma = I$ and $\Omega = \sigma^2(X^T X)^{-1}$ we obtain,

$$\hat{R}_{\beta,h} = \|e_h\|^2 + 2\sigma^2\left(\text{trace}\left(V \frac{\partial \hat{\beta}_h}{\partial z^T}\right)\right) \tag{6.12}$$

where $V = (X^T X)^{-1}$. Note that as with all SURE formulae we can have $\hat{\beta}_h$ to be an arbitrary nonlinear function of z.

6.2 l_1 Penalized Least Squares with Grouped Variables

6.2.1 Group LASSO Preview

This section gives an overview of the group LASSO problem [278]. Consider the case similar to the network topology identification problem discussed in chapter 5, where elements of β vector are partitioned as $\beta = (\beta_1^T, \ldots, \beta_g^T)^T$, into g groups with dimension $\dim(\beta_u) = p_u$. The X matrix has the corresponding partitions $X = [X_{(1)}, \ldots, X_{(g)}]$. It is assumed that the regressors are centered and column scaled such that $X_{(u)}^T X_{(u)} = I_{p_u}$. Group LASSO estimates β by minimizing the following penalized least squares criterion discussed in section 5.2,

$$J(\beta) = \frac{1}{2}\|y - \sum_1^g X_{(u)}\beta_u\|^2 + h\sum_1^g \|\beta_u\| \tag{6.13}$$

Kuhn-Tucker conditions for the optimal solution of (6.13) for $1 \leq u \leq g$ are [6], [278],

$$\begin{aligned} -X_{(u)}^T(y - X\beta) + \frac{h\beta_u\sqrt{p_u}}{\|\beta_u\|} = 0, \qquad & \|\beta_u\| \neq 0 \\ \| - X_{(u)}^T(y - X\beta)\| \leq h\sqrt{p_u}, \qquad & \|\beta_u\| = 0 \end{aligned}$$

and this leads to the solution,

$$\hat{\beta}_{h,u} = (1 - \alpha_u)H_u S_u, 1 \leq u \leq g, \tag{6.14}$$

where $H_u = H(\|S_u\| - h\sqrt{p_u})$, $H(\cdot)$ is the Heaviside step function, $\alpha_u = \frac{h\sqrt{p_u}}{\|S_u\|}$ and,

$$S_u = X_{(u)}^T(y - X\hat{\beta}_h + X_{(u)}\hat{\beta}_{h,u}) = X_{(u)}^T(y - X\hat{\beta}_h) + \hat{\beta}_{h,u} \tag{6.15}$$

(6.14) must be solved iteratively as in [278] by cyclic descent.

6.2.2 Derivation of SURE for $R_{\beta,h}$ for the Group LASSO

From (6.12) we can see that we need to find trace $\left(V\partial\hat{\beta}_h/\partial z^T\right)$ but there is apparently a problem because from (6.14) and (6.15) it appears that $\hat{\beta}_h$ is a function of y not z. Fortunately this problem can be overcome with the following device. Add and subtract Xz inside (6.15) to get,

$$\begin{aligned} S_u &= X_{(u)}^T(y - Xz + Xz - X\hat{\beta}_h) + \hat{\beta}_{h,u} \\ &= X_{(u)}^T(Xz - X\hat{\beta}_h) + \hat{\beta}_{h,u} \end{aligned} \tag{6.16}$$

since z is the ordinary least squares estimator, $X_{(u)}^T(y - Xz) = 0$. Now (6.14) becomes an implicit equation whose solution $\hat{\beta}_{h,u}$ will be a function of z as required. Since (6.14) gives the solution to (6.13) using the chain rule we can write,

$$\frac{\partial\hat{\beta}_{h,u}}{\partial z^T} = \frac{\partial\hat{\beta}_{h,u}}{\partial S_u^T}\,\frac{\partial S_u}{\partial z^T} \tag{6.17}$$

From (6.16) we can derive,

$$\frac{\partial S_u}{\partial z^T} = X_{(u)}^T X - X_{(u)}^T X\frac{\partial\hat{\beta}_h}{\partial z^T} + \frac{\partial\hat{\beta}_{h,u}}{\partial z^T} \tag{6.18}$$

Using Dirac delta function $\delta(\cdot)$, define $\delta_u = \delta(\|S_u\| - h\sqrt{(}p_u))$, then from (6.14),

$$\begin{aligned} \frac{\partial\hat{\beta}_{h,u}}{\partial S_u^T} =& (1-\alpha_u)H_u I_{p_u} + (1-\alpha_u)\delta_u\frac{d\|S_u\|}{dS_u^T}S_u - h\sqrt{p_u}\frac{d1/\|S_u\|}{dS_u^T}H_u S_u \\ =& (1-\alpha_u)H_u I_{p_u} + (1-\alpha_u)\delta_u\hat{S}_u S_u^T + \alpha_u H_u\hat{S}_u\hat{S}_u^T \end{aligned} \tag{6.19}$$

where

$$\hat{S}_u = \frac{d\|S_u\|}{dS_u^T} = \frac{d(S_u^T S_u)^{1/2}}{dS_u^T} = \frac{S_u}{\|S_u\|},$$

Note that $\hat{S}_u$ is a unit vector. The second term of (6.19) vanishes and we obtain,

$$\frac{\partial \hat{\beta}_{h,u}}{\partial S_u^T} = H_u[(1-\alpha_u)I_{p_u} + \alpha_u \hat{S}_u \hat{S}_u^T] \tag{6.20}$$

Substituting (6.18) and (6.20) in (6.17) for $1 \leq u \leq g$ we get,

$$\frac{\partial \hat{\beta}_{h,u}}{\partial z^T} = H_u[(1-\alpha_u)I_{p_u} + \alpha_u \hat{S}_u \hat{S}_u^T] \times \left[X_{(u)}^T X - X_{(u)}^T X \frac{\partial \hat{\beta}_h}{\partial z^T} + \frac{\partial \hat{\beta}_{h,u}}{\partial z^T} \right] \tag{6.21}$$

Introduce $D = \text{block diagonal } (\Delta_u)$ with $\Delta_u = H_u I_{p_u}(1-\alpha_u) + H_u \alpha_u \hat{S}_u \hat{S}_u^T$. Then we can rewrite this as,

$$\begin{aligned} \frac{\partial \hat{\beta}_h}{\partial z^T} &= D \left[X^T X - X^T X \frac{\partial \hat{\beta}_h}{\partial z^T} + \frac{\partial \hat{\beta}_h}{\partial z^T} \right] \\ &= \left[I - D + DX^T X \right]^{-1} DX^T X. \end{aligned} \tag{6.22}$$

Thus,

$$\begin{aligned} \text{trace}\left(V \frac{\partial \hat{\beta}_h}{\partial z^T} \right) &= \text{trace}\left(V \left[I - D + DX^T X \right]^{-1} DX^T X \right), \\ &= \text{trace}\left(\left[I - D + DX^T X \right]^{-1} DX^T XV \right), \\ &= \text{trace}\left(\left[I - D + DX^T X \right]^{-1} D \right) = \tau_h, \end{aligned}$$

since $X^T XV = I$. From (6.12) SURE for $R_{\beta,h}$ for group LASSO now becomes,

$$\hat{R}_{\beta,h} = \|e_h\|^2 + 2\sigma^2 \tau_h. \tag{6.23}$$

This expression can actually be simplified further. Introduce $\Gamma_0 = \{u : \hat{\beta}_u = 0\} = \{u : H_u = 0\}$, $\Gamma_c = \{u : \hat{\beta}_u \neq 0\} = \{u : H_u = 1\}$. Denote $g_h = \dim(\Gamma_c) = \sharp$ groups with non-zero parameters. Then reorder the entries of D such that the

indices $u \in \Gamma_c$ come first as follows,

$$D = \begin{pmatrix} D_{c_h} & 0 \\ 0 & 0_{(p-c_h)\times(p-c_h)} \end{pmatrix},$$

where D_{c_h} is a block diagonal matrix of dimension $c_h \times c_h$, $c_h = \sum_1^{g_h} p_u$ and $p = \sum_1^g p_u$. Thus,

$$I - D + DX^TX = \begin{pmatrix} I_{c_h} - D_{c_h} & 0 \\ 0 & I_{p-c_h} \end{pmatrix} + \begin{pmatrix} D_{c_h} & 0 \\ 0 & 0_{(p-c_h)\times(p-c_h)} \end{pmatrix} \begin{pmatrix} X_c^TX_c & X_c^TX_0 \\ X_0^TX_c & X_0^TX_0 \end{pmatrix}$$

where the columns of X have been reordered and then partitioned it to conform with D. Continuing,

$$I-D+DX^TX = \begin{pmatrix} I_{c_h} - D_{c_h} + D_{c_h}X_c^TX_c & D_{c_h}X_c^TX_0 \\ 0 & I_{p-c_h} \end{pmatrix}$$

The inverse of this block triangular matrix is easily seen to be,

$$(I-D+DX^TX)^{-1} = \begin{pmatrix} M_{c_h}^{-1} & -M_{c_h}^{-1}D_{c_h}X_c^TX_0 \\ 0 & I_{p-c_h} \end{pmatrix}$$

where $M_{c_h} = I_{c_h} - D_{c_h} + D_{c_h}X_c^TX_c$. Thus,

$$\begin{aligned} \tau_h &= \text{trace}\left(\begin{pmatrix} M_{c_h}^{-1} & -M_{c_h}^{-1}D_{c_h}X_c^TX_0 \\ 0 & I_{p-c_h} \end{pmatrix} \begin{pmatrix} D_{c_h} & 0 \\ 0 & 0 \end{pmatrix}\right) \\ &= \text{trace}(M_{c_h}^{-1}D_{c_h}) \end{aligned}$$

To simplify further consider that,

$$\begin{aligned} \Delta_u &= H_u(I_{p_u}(1-\alpha_u) + \alpha_u\hat{S}_u\hat{S}_u^T) \\ &= H_u[(1-\alpha_u)(I - \hat{S}_u\hat{S}_u^T) + \hat{S}_u\hat{S}_u^T] \end{aligned}$$

Since $I - \hat{S}_u\hat{S}_u^T$ and $\hat{S}_u\hat{S}_u^T$ are orthogonal to each other, we deduce that, $\Delta_u^{\frac{1}{2}} = H_u[\sqrt{(1-\alpha_u)}(I - \hat{S}_u\hat{S}_u^T) + \hat{S}_u\hat{S}_u^T]$. Now define $D_{c_h}^{\frac{1}{2}} = \text{blockdiag } \Delta_u^{\frac{1}{2}}$. Using this

we can write,

$$\begin{aligned}\tau_h &= \text{trace}(M_{c_h}^{-1} D_{c_h}) \\ &= \text{trace}((D_{c_h}^{\frac{1}{2}}(I_{c_h} - D_{c_h} + D_{c_h}^{\frac{1}{2}} X_c^T X_c D_{c_h}^{\frac{1}{2}}) D_{c_h}^{-\frac{1}{2}})^{-1} D_{c_h}) \\ &= \text{trace}(\bar{M}_{c_h}^{-1} D_{c_h}) \end{aligned} \tag{6.24}$$

where

$$\bar{M}_{c_h} = I_{c_h} - D_{c_h} + D_{c_h}^{\frac{1}{2}} X_c^T X_c D_{c_h}^{\frac{1}{2}}. \tag{6.25}$$

Furthermore,

$$\begin{aligned} X_k D_{c_h}^{\frac{1}{2}} &= [X_{(1)}, \ldots, X_{(c_h)}] \text{blockdiag} \Delta_u^{\frac{1}{2}}, \\ &= [X_{(1)} \Delta_1^{\frac{1}{2}}, \ldots, X_{(c_h)} \Delta_{c_h}^{\frac{1}{2}}]. \end{aligned}$$

The block diagonal entries of $\bar{M}_{c_h}$ are thus, $\bar{M}_{uu} = I - \Delta_u + \Delta_u^{\frac{1}{2}} X_{(u)}^T X_{(u)} \Delta_u^{\frac{1}{2}} = I - \Delta_u + \Delta_u = I$. And off block diagonal entries of $\bar{M}_{c_h}$ are, $\bar{M}_{uv} = [\Delta_u^{\frac{1}{2}} X_{(u)}^T X_{(v)} \Delta_v^{\frac{1}{2}}]$. Putting all this together we first compute $\bar{M}_{c_h}^{-1} = [N_{uv}]$. Then from (6.24), $\tau_h = \text{trace}(\bar{M}_{c_h}^{-1} D_{c_h}) = \sum_1^{g_h} \text{trace}(N_{uu} \Delta_u)$. Since $H_u = 1, u \in \Gamma_c, \Delta_u = (1 - \alpha_u)(I - \hat{S}_u \hat{S}_u^T) + \hat{S}_u \hat{S}_u^T$. Thus,

$$\tau_h = \sum_1^{g_h} (\text{trace}(N_{uu})(1 - \alpha_u) + \alpha_u \hat{S}_u^T N_{uu} \hat{S}_u) \tag{6.26}$$

Collecting these results together we have:

Result I: Provided $\bar{M}_{c_h}$ is invertible, SURE for $R_{\beta,h}$ for group LASSO is given by $\hat{R}_{\beta,h} = \|e_h\|^2 + 2\sigma^2 \tau_h$, $e_h = z - \hat{\beta}_h = (X^T X)^{-1} X^T y - \hat{\beta}_h$ where $\hat{\beta}_h$ is obtained by solving (6.14) by cyclic decent. Also τ_h is given in (6.26) where $\hat{S}_u = S_u / \|S_u\|$ and S_u is defined in (6.15). Further $[N_{uv}] = \bar{M}_{c_h}^{-1}$ while $\bar{M}_{ch}$ is defined in (6.25). Note that the required computations are modest.

Note the special cases:

(i) $\underline{p_u = 1}$. Then $\tau_h = \sum_1^{g_h} N_{uu}(1 - \alpha_u + \alpha_u) = \sum_1^{g_h} N_{uu}$. Thus $\hat{R}_{\beta,h} = \|e_h\|^2 + 2\sigma^2 \sum_1^{g_h} N_{uu}$.

(ii) $\underline{X^T X = I}$. Then $M_{uv} = 0, u \neq v$, thus $\bar{M}_{c_h} = I \Rightarrow N_{uu} = I_{p_u} \Rightarrow \tau_h = \sum_1^{g_h} p_u(1 - \alpha_u) + \alpha_u$.

For comparison SURE for $R_{\mu,h}$ for group LASSO [223] with $K = I$ is given by,

$$\hat{R}_{\mu,h} = \| e \|^2 + 2\sigma^2(c_h - \rho_h), \tag{6.27}$$

where $e = y - \hat{\mu}_h$ and,

$$\rho_h = \sum_1^{g_h} (\text{trace}(N_{uu})\alpha_u - \alpha_u \hat{S}_u^T N_{uu} \hat{S}_u). \tag{6.28}$$

Conditions to ensure the invertibility of $\bar{M}_{c_h}$ is given in [223].

6.3 Group LASSO Model Selection Simulation

Topology identification of a sparsely connected network of dynamic systems was introduced in chapter 5. As discussed in section 5.1 recovering the topology of a network using time series data inherently involves sparse regression of grouped variables. In this section data is generated by simulating a network of dynamic systems as given in section 5.1 and the topology is estimated by optimizing (6.13) using group LASSO. Here a finite impulse response (FIR) model is used for for model expansion. Similar results can be obtained with Laguerre basis functions.

The performance of h_β the minimizer of $\hat{R}_{\beta,h}$ is compared with with h_μ the minimizer of $\hat{R}_{\mu,h}$ and h_{BIC} the minimizer of BIC in estimating h_b the minimizer of $R_{\beta,h}$.

Usually BIC is used to select discrete tuning parameters. However, as mentioned in previous chapters, the value of h determines the number of non zero coefficients retained by the algorithm, thus enabling us to use BIC to select h.

A network with 10 nodes was simulated and each node was connected to 6 randomly selected nodes. The order of all autoregressive models were set to 5 and the system parameters (β) were drawn from a normal distribution with 0 mean and 0.04 variance. Then the location of the poles of the overall network were

analyzed to ensure stability. The network was excited with Gaussian random noise with 0 mean and 0.01 variance and $n = 300$ time samples recorded at each node which makes up the y^i vectors.

The matrix X in (5.5e), (5.5f) was generated and then the columns were centered such that the mean is 0. $H_{i,j}$ partition of X was orthonormalized ($H_{i,j}^T H_{i,j} = I_m, j = 1, \ldots, N$). For a selected stable network 500 y^i vectors were generated and the coefficient vector ($\hat{\beta}_h^i$) was estimated using group LASSO at each node for a range of h values.

$R_{\beta,h}$, $R_{\mu,h}$, $\hat{R}_{\beta,h}$ (6.23), $\hat{R}_{\mu,h}$ (6.27), and BIC [200, 135, 123] were calculated for each estimate and their corresponding minimizers h_b, h_m, h_β, h_μ, h_{BIC} were found. Note when calculating $R_{\beta,h}$ and $R_{\mu,h}$ the original system parameters (β) have to be scaled to overcome the effects of the orthonormalization of the $H_{i,j}$ partitions. Since group LASSO is applied at each node, only the results of one node will be shown here. The results of other nodes are similar.

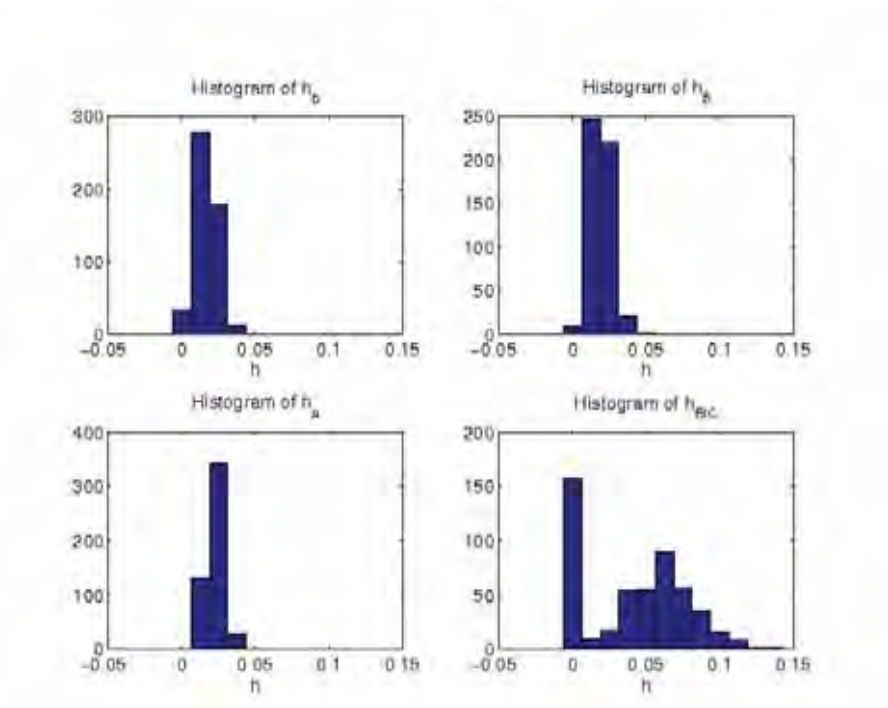

Figure 6.1: Histograms of h_b, h_β, h_μ and h_{BIC} when estimating the connectivity of a single node over 500 realizations of a network with 10 nodes with 6 connections per node.

The histograms of the h that minimized the three model selection criteria is compared with that of h_b in figure 6.1. The histograms of h_b and h_β follow the same pattern and they are centered around the same region. The histogram of h_β has slightly wider distribution than that of h_b. This is expected due to the noise in the system. The histogram of h_μ does not have the same center as that of h_b. The histogram of h_{BIC} has a very wide distribution and most of the minimum has occurred at $h = 0$, thus in this application BIC is unfavorable. The relative deviation was calculated as follows,

$$\text{Deviation}_h = \text{D}_h = \frac{\text{median}(h) - \text{median}(h_b)}{\text{median}(h_b)},$$

where h is one of h_β, h_μ, h_{BIC}. The deviations were found to be $D_{h_\beta} = 0$, $D_{h_\mu} = 0.9804$ and $D_{h_{BIC}} = 2.9412$. Thus h_β is a better estimate of h_b.

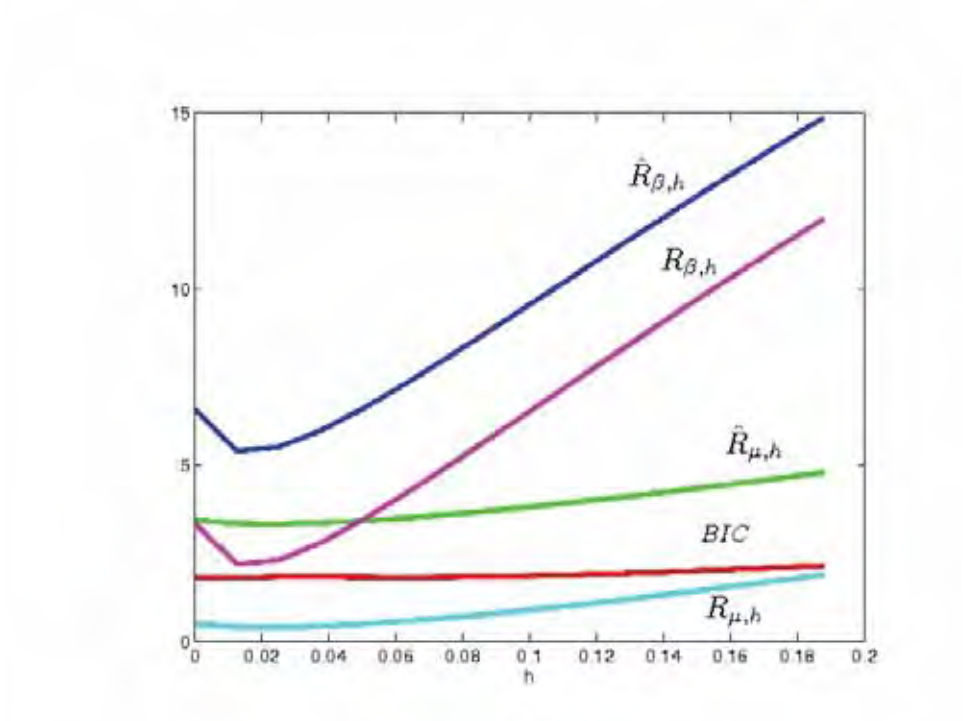

Figure 6.2: Model selection criterion as a function of h when estimating the transfer function of a single node within a network of 10 nodes with 6 connections per node, averaged over 500 replications.

The functional dependance of the model selection criterion with h averaged

over 500 replications is given in figure 6.2. The functional dependance of $R_{\beta,h}$ and $R_{\mu,h}$ are also displayed for comparison. The value of $\hat{R}_{\beta,h}$ is different from that of $R_{\beta,h}$ because the terms that did not depend on $\hat{\beta}_h$ were dropped in the derivation of $\hat{R}_{\beta,h}$. However we can clearly see that $\hat{R}_{\beta,h}$ follows the same pattern as $R_{\beta,h}$. The minimum of $R_{\beta,h}$ and $\hat{R}_{\beta,h}$ were reached at 0.0128 and the minimum of $R_{\mu,h}$ and $\hat{R}_{\mu,h}$ were reached at 0.0253. We can clearly see that h_m is different from h_b. Further more h_β is superior to the minimizers of other model selection criterion when estimating h_b. From figure 6.1 and 6.2 we can see that in this example BIC fails in selecting an appropriate h.

6.4 Conclusion

This chapter derived a general expression of SURE for MSE_β in a regression problem with possibly non linear estimation of the regression coefficients. Previous work gave SURE for MSE_μ. This result was then applied to obtain SURE for MSE_β for group LASSO. Simulation results for a network topology problem show the accuracy of the model selector as well as its superiority when compared to other model selection methods such as SURE for MSE_μ and BIC in obtaining an estimate with minimum MSE_β. In this application BIC in particular exhibits very poor behavior.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The aim of this thesis was twofold. First it extended the understanding of the l_0 norm in sparse signal processing by unmasking some myths, developing novel algorithms and by comparing them with alternative sparse algorithms. It also addressed some issues of model selection.

Chapter 1 outlined the wide range of applications which motivated the development of sparse signal processing and gave an overview of the available sparse modeling algorithms. It presented a detailed discussion about sparsity measures and their properties. The l_0 norm is severely criticized in the literature as being unsuitable for sparse signal processing. These criticisms were rebutted in this chapter.

Novel sparse algorithms were developed in chapters 2 and 3. Chapter 2 was based on direct minimization of the l_0 norm (l_0 denoising). First it investigated the ability of existing algorithms to perform exact l_0 denoising. Then it developed the L0LS-CD algorithm and its multivariate (V-L0LS-CD) and group (gL0LS-CD) regression variants. The L0LS-CD algorithm minimizes the l_0 penalized least squares via cyclic descent. Non-trivial stability analysis of the L0LS-CD algorithm was developed. Issues of computational speed and initialization were addressed and the impact of penalty parameter on the performance of the algorithm was illustrated. Simulation results showed that L0LS-CD and V-L0LS-CD outperformed other existing algorithms.

Chapter 3 was based on smooth approximations of sparsity measures including those of the l_0 norm. The QC algorithm developed in chapter 3 can minimize the least squares criterion penalized with any penalty with the quadratic concave property via the majorization minimization (MM) technique. Since the common smooth approximations of the l_0 norm are quadratic concave functions, the QC algorithm is applicable. This chapter presents the convergence analysis of the QC algorithm. An interesting comparison shows that QC minimizes the criterion even when the Newton algorithm fails. Simulation results showed that when many singular values of the regression matrix are close to zero, the QC algorithm outperformed L0LS-CD and pIHT algorithms.

Chapters 4 and 5 were based on applications of sparse modeling. Chapter 4 developed the SCSI algorithm which is a sparse transfer function estimation method for systems with coloured noise. Laguerre basis functions were used for model expansion which guarantees stability unlike existing methods. SCSI simulations further confirms that L0LS-CD outperforms L1LS.

Chapter 5 was based on topology identification of sparsely connected networks of dynamic systems. The transfer functions were expanded using Laguerre basis functions and gL0LS-CD was used for topology identification. Simulation results shows that gL0LS-CD outperforms group LASSO.

Although sparse signal processing has attracted a lot of attention, the importance of model selection has been widely neglected. Most model selection criterion are based on the signal or prediction mean squared error. Chapter 6 develops an SURE criterion instead for parameter mean square error and obtains the estimator for l_1 penalized least squares problem with grouped variables. Simulations based on a network topology problem showed the accuracy of this method as well as its superiority compered to other methods.

In conclusion this thesis has made a significant contribution to the area of sparse signal processing and model selection. Future avenues for research related to the topics discussed in this thesis are listed below.

7.2 Future Work

1. Chapter 2: l_0 denoising.
    - L0LS-CD, V-L0LS-CD and gL0LS-CD were developed for systems with white noise. These methods can be extended to handle coloured noise. SCSI-L0, developed in chapter 4 is a special instance of this in the context of system identification. A general method to handle coloured noise can be developed by following a similar principle.
    - So far SURE criteria have been developed for the l_1 norm. SURE for MSE_μ and MSE_β for L0LS can be developed.
    - This thesis presents the stability analysis of L0LS-CD on over-determined systems. This analysis can be extended to under-determined systems.
    - So far L0LS-CD has been applied to the application of sparse transfer function estimation and gL0LS-CD has been applied to the applications of magnetoencephalography and network topology identification. It would be interesting to analyze the performance of L0LS-CD, V-L0LS-CD and gL0LS-CD in the context of other sparse modeling applications.
    - In [30] gL0LS-CD is applied to individual time samples separately. It would be interesting to extend gL0LS-CD to incorporate temporal correlation.
2. Chapter 3: QC algorithm.
    - The QC algorithm can also be extended to handle coloured noise.
    - The QC algorithm is more general and has a wider scope than presented in chapter 3. QC has been extended to handle noisy independent component analysis (ICA) in [222].
    - The stability analysis of the QC algorithm can be extended to support under-determined systems.
    - So far QC is applied to each time sample of noisy ICA separately. It would be interesting to extend this method to introduce temporal continuity.

3. Chapter 4: Application of sparse modeling - transfer function estimation.

    - SCSI algorithm estimates a stable, sparse transfer function. However the estimated noise model is non-sparse. SCSI algorithm can be extended to enforce sparsity to the noise model, however this would require the noise model to be changed because stability cannot be guaranteed if sparsity is enforced on an autoregressive model.
    - Many tuning parameters have to be selected for the SCSI algorithm. In chapter 4 these were selected using BIC. SURE criterion can be developed for this purpose.

4. Chapter 5: Application of sparse modeling - network topology identification.

    - Stability of a multiple input multiple output system is quite complicated specially when sparsity is enforced. Although the method presented in chapter 5 can successfully identify the topology of the network it cannot guarantee the stability of the estimated network. Developing a method that can enforce sparsity on to the network links, as well as to the transfer functions of individual links, while guaranteeing stability will be an interesting avenue for future work.

5. Chapter 6: Parameter Based Model Selection via SURE.

    - Although SURE is superior to the other model selection criteria it was not used in the simulations of chapters 2-5 because it needs further development. So far the SURE criterion has only been developed for the l_1 penalized least squares problem. Thus SURE can be extended to handle other sparsity measures.

References

[1] R. Adams, Y. Xu, and F. Canning. Sparse pseudo inverse of the discrete plane wave transform. *IEEE Transactions on Antennas and Propagation*, 56:475–484, 2008. 2

[2] M. Aharon, M. Elad, and A. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear Algebra and its Applications*, 416:4867, 2006. 4

[3] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54:4311–4322, 2006. 4

[4] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proc. IEEE Second International Symposium on Information Theory*, pages 267–281, 1973. 19

[5] H. Akcay and B. Ninness. Rational basis functions for robust identification from frequency and time-domain measurements. *Automatica*, 34:1101–1117, 1998. 89

[6] S. Alliney and S. A. Ruzinsky. An algorithm for the minimization of mixed l_1 and l_2 norms, with application to bayesian estimation. *IEEE Transactions on Signal Processing*, 42:618 – 627, 1994. 14, 28, 95, 106, 115

[7] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, New York, 1958. 17

[8] A. Antoniadis and J. Fan. Regularization of wavelet approximation. *Journal of the American Statistical Association*, 96:939–967, 2001. 14

[9] T. Apostol. *Mathematical Analysis*. Addison-Wesley, New York, 1974. 85

[10] A. Arnold, Y. Liu, and N. Abe. Estimating brain functional connectivity with sparse multivariate autoregression. In *Proc. 13th ACM International Conference on Knowledge Discovery and Data Mining*, pages 66–75, San Jose, California, USA, 2007. 102

[11] M. Ayazoglu, M. Sznaier, and N. Ozay. Blind identification of sparse dynamic networks and applications. In *Proc. IEEE Conference on Decision and Control*, pages 2944–2950, Orlando, FL, USA, 2011. 101, 102

[12] J. S. Bailly, P. Monestiez, and P. Lagacherie. Modelling spatial variability along drainage networks with geostatistics. *Mathematical Geology*, 38:515–539, 2006. 101

[13] R. Baraniuk and P. Steeghs. Compressive radar imaging. In *Proc. IEEE Radar Conference*, pages 128–133, Waltham, MA, U.S.A., 2007. 2

[14] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New Jersey, 1980. 3

[15] H. J. W. Belt and A. C. Den Brinker. Optimality condition for truncated generalized laguerre networks. *International Journal of Circuit Theory and Applications*, 23:227–235, 1995. 90

[16] C. R. Berger, J. Areta, K. Pattipati, and P. Willett. Compressed sensing - a look beyond linear programming. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3857–3860, Las Vegas, Nevada, U.S.A., 2008. 16

[17] J. Bi, K. P. Bennett, M. Embrechts, C. M. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *The Journal of Machine Learning Research*, 3:1229–1243, 2003. 2

[18] T. Blumensath. Accelerated iterative hard thresholding. *Signal Processing*, 92:752–756, 2009. 17

[19] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27:265–274, 2009. 17

[20] T. Blumensath and M.E. Davies. Iterative thresholding for sparse approximation. *Journal of Fourier Analysis and Application*, 14:629–654, 2008. 16, 28, 49, 52, 53, 56, 71

[21] T. Blumensath and M.E. Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4:298–309, 2010. 52, 53

[22] P. Bodin and B. Wahlberg. Thresholding in high order transfer function estimation. In *Proc. IEEE Conference on Decision and Control*, pages 3400–3405, Florida, USA, 1994. 88, 89

[23] A. Bolstad, B. D. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing*, 59:2628–2641, 2011. 18, 101, 102, 107, 109

[24] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 12, 14

[25] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373384, 1995. 14

[26] S. P. Brooks, N. Friel, and R. King. Classical model selection via simulated annealing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:503–520, 2003. 12

[27] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006. 2

[28] E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008. 16, 73

[29] B. Cassidy, C. J. Long, C. Rae, and V. Solo. Identifying fMRI model violations with lagrange multiplier tests. *IEEE Transactions on Medical Imaging*, 31:1481–1492, 2012. 89

[30] B. Cassidy, V. Solo, and A. J. Seneviratne. Grouped L0 least squares penalized magnetoencephalography. In *Proc. IEEE International Symposium on Biomedical Imaging*, pages 868–871, Barcelona, Spain, 2012. 126

[31] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14:707–710, 2007. 15

[32] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24:1–14, 2008. 15

[33] J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008. 51

[34] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Transactions on Signal Processing*, 54:4634–4643, 2006. 18

[35] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50:18731896, 1989. 13

[36] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001. 4, 14

[37] X. Chen, Z. J. Wang, and M. J. McKeown. fMRI group studies of brain connectivity via a group robust lasso. In *Proc. IEEE International Conference on Image Processing*, pages 589–592, Hong Kong, China, 2010. 18

[38] H. Chung, K. Lee, and J. Koo. A note on bootstrap model selection criterion. *Statistics and Probability Letters*, 26:35–41, 1996. 19

[39] P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4:1168–1200, 2005. 71, 77, 85, 86

[40] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 53:2477–2488, 2005. 17, 39, 50, 60

[41] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16:2080–2095, 2007. 2

[42] E. DallAnese, J.-A. Bazerque, H. Zhu, and G. B. Giannakis. Group sparse total least-squares for cognitive spectrum sensing. In *Proc. IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications*, pages 96 –100, San Francisco, California, U.S.A., 2011. 18

[43] H. Dalton. The measurement of the inequity of incomes. *Economic Journal*, 30:348–361, 1920. 7, 8

[44] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. 10

[45] J. Doak. An evaluation of feature selection methods and their application to computer security. *University of California at Davis, Tech Report CSE-92-18*, 1992. 12

[46] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20:1838–1857, 2011. 2

[47] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006. 2

[48] D. L. Donoho. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Communications On Pure and Applied Mathematics*, 59:907–934, 2006. 14

[49] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. In *Proc. National Academy of Sciences of the United States of America*, pages 2197–2202, 2003. 14

[50] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001. 14

[51] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B*, 54:41–81, 1992. 21

[52] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 58:1094–1121, 2012. 13

[53] N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, 1966. 12

[54] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979. 19

[55] B. Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99:619–642, 2004. 19

[56] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004. 14, 112

[57] M. A. Efroymson. Stepwise regression-a backward and forward look. In *Proc. Eastern Regional Meetings of the Institute of Mathematical Statistics*, 1966. 12

[58] M. Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137:334–353, 2007. 101

[59] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15:3736–3745, 2006. 2, 4

[60] M. Elad, M.A.T. Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98:972–982, 2010. 2

[61] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57:471–481, 2009. 21, 111

[62] Y. C. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Transactions on Information Theory*, 56:505–519, 2010. 18

[63] Y.C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58:3042–3054, 2010. 18, 41

[64] Y.C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55:5302–5316, 2009. 18, 41

[65] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, page 24432446, Phoenix, Arizona, USA, 1999. 4

[66] K. Engan, B. D. Rao, and K. Kreutz-Delgado. Frame design using FOCUSS with method of optimal directions (MOD). In *Proc. Norwegian Signal Processing Symposium*, page 6569, Sem Gjestegård, Asker, Norway, 1999. 4

[67] J. Fan. Comment on wavelets in statistics: A review by a. antoniadis. *Italian Journal of Statistics*, 6:97–144, 1997. 14

[68] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001. 14

[69] I. J. Fevrier, S. B. Gelfand, and M. P. Fitz. Reduced complexity decision feedback equalization for multipath channels with large delay spreads. *IEEE Transactions on Communications*, 47:927–937, 1999. 17

[70] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE journal of selected topics in signal processing (Special Issue on Convex Optimization Methods for Signal Processing)*, 1:586–598, 2007. 14

[71] S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26:395–407, 2009. 15

[72] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148, 1993. 15

[73] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302–332, 2007. 14, 32, 36

[74] W. J. Fu. Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998. 14, 15, 32, 36, 52

[75] T. S. Gardner, S. Shimer, and J. J. Collins. Inferring microbial genetic networks. *ASM News*, 70:121–126, 2004. 18, 101

[76] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57:4686–4698, 2009. 15, 70, 71

[77] D. Geeman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:367–383, 1992. 71

[78] D. Geeman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:932–946, 1995. 71

[79] E. I. George and D. P. Foster. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22:1947–1975, 1994. 20

[80] E. I. George and D. P. Foster. Calibration and empirical bayes variable selection. *Biometrika*, 87:731–747, 2000. 20

[81] P. Georgiev, F. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16:992–996, 2005. 2

[82] W. Gersch. Spectral analysis of EEG's by autoregressive decomposition of time series. *Mathematical Biosciences*, 7:205–222, 1970. 101

[83] W. Gersch. Causality or driving in electrophysiological signal analysis. *Mathematical Biosciences*, 14:177–196, 1972. 101

[84] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen. Sparse linear prediction and its applications to speech processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:1644–1657, 2012. 2

[85] C. Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126, 1921. 7

[86] D.E. Goldberg. *Genetic Algorithm in Search, Optimization, and Machine Learning*. Addison Wesley, 1989. 12

[87] D.E. Goldberg. Genetic and evolutionary algorithms come of age. *Communications of the ACM*, 37:113–119, 1994. 12

[88] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979. 19

[89] I. F. Gorodnitsky, J. S. George, and B. D. Rao. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. *Electroencephalography and Clinical Neurophysiology*, 95:231–251, 1995. 17

[90] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45:600 – 616, 1997. 15, 52

[91] I.F. Gorodnitsky and B.D. Rao. A recursive weighted minimum norm algorithm: Analysis and applications. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 456–459 vol.3, Minneapolis, Minnesota, U.S.A., 1993. 15

[92] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969. 101

[93] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms. *The Journal of Fourier Analysis and Applications*, 14:655–687, 2008. 17

[94] A. Gupta, G. Karypis, and V. Kumar. Highly scalable parallel algorithms for sparse matrix factorization. *IEEE Transactions on Parallel and Distributed Systems*, 8:502–520, 1997. 2

[95] Y. Hamamoto, S. Uchimura, Y. Matsuura, T. Kanaoka, and S. Tomita. Evaluation of the branch and bound algorithm for feature selection. *Pattern Recognition Letters*, 11:453–456, 1990. 12

[96] M. Hanke. Limitations of the l-curve method in ill-posed problems. *BIT Numerical Mathematics*, 36:287–301, 1996. 19, 50

[97] P. C. Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review*, 34:561–580, 1992. 19

[98] P. C. Hansen and D. P. O'Leary. The use of the l-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14:1487–1503, 1993. 19

[99] S. Haufe, K. Müller, G. Nolte, and N. Krämer. Sparse causal discovery in multivariate time series. In *Proc. NIPS workshop on causality*, pages 97–106, Westin Hilton, BC, Canada, 2008. 101, 102

[100] K. K. Herrity, A. C. Gilbert, and J. A. Tropp. Sparse approximation via iterative thresholding. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 624–627, Toulouse, France, 2006. 16

[101] P. Heuberger, P. Van den Hof, and O. Bosgra. Modelling linear dynamical systems through generalized orthonormal basis functions. In *Proc. 12th ZFAC World Congress*, pages 283–286, Sydney, Australia, 1993. 89

[102] P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra. A generalized orthonormal basis for linear dynamical systems. *IEEE Transactions Automatic Control*, 40:451–465, 1995. 89

[103] A.E. Hoerl and R.W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970. 15

[104] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970. 15

[105] J. A. Hogbom. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astronomy and Astrophysics Supplement*, 15:417–426, 1974. 12, 13

[106] J. H. Holland. *Adaptation in Natural and Artificial Systems*. Ann Arbor MI University of Michigan Press, 1975. 12

[107] R. Horst and N. V. Thoai. DC programming: Overview. *Journal of Optimization Theory and Applications*, 103:1–41, 1999. 70

[108] J. Huang, J. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36:587–613, 2008. 16

[109] P. Huber. *Robust Statistics*. J Wiley, New York, 1981. 77

[110] H. M. Hudson and T. C. M. Lee. Maximum likelihood restoration and choice of smoothing parameter in deconvolution of image data subject to poisson noise. *Computational Statistics and Data Analysis*, 26:393–410, 1998. 21

[111] D. Hunter and R. Li. Variable selection using MM algorithms. *Annals of Statistics*, 33:1617–1642, 2005. 80

[112] D. R. Hunter and K. Lange. Rejoinder to discussion of optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9:52–59, 2000. 71

[113] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30–37, 2004. 71, 75, 76

[114] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55:4723–4741, 2009. 8, 9, 11

[115] M. Hyder and K. Mahata. An approxmate l0 norm minimization algorithm for compressed sensing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3365–3368, Taipei, Taiwan, 2009. 16, 52

[116] M. Hyder and K. Mahata. Direction-of-arrival estimation using a mixed $l_{2,0}$ norm approximation. *IEEE Transactions on Signal Processing*, 58:4646–4655, 2010. 16, 17, 39, 50, 52, 60

[117] M. Hyder and K. Mahata. An improved smoothed l^0 approximation algorithm for sparse representation. *IEEE Transactions on Signal Processing*, 58:2194–2205, 2010. 16, 52

[118] B. D. Jeffs. Sparse inverse solution methods for signal and image processing applications. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1885–1888, Seattle, Washington, U.S.A., 1998. 17

[119] J. Karvanen and A. Cichocki. Measuring sparseness of noisy signals. In *Proc. International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 125–130, Nara, Japan, 2003. 8, 9, 11

[120] R. E. Kass and L. Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90:928–934, 1995. 20

[121] Y. Kinoshita and Y. Ohta. Continuous-time system identification using compactly-supported filter kernels generated from Laguerre basis functions. In *Proc. IEEE Conference on Decision and Control*, pages 4461–4466, Atlanta, GA, 2010. 89

[122] S. Kirkpatrick, C.D. Gellat Jr., and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983. 12

[123] S. Konishi and G. Kitagawa. *Information criteria and statistical modeling.* Springer, 2008. 50, 121

[124] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15:349–396, 2003. 4

[125] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 57:425–437, 1995. 43

[126] R. M. Leahy and B. D. Jeffs. On the design of maximally sparse beamforming arrays. *IEEE Transactions on Antennas and Propagation*, 39:1178–1187, 1991. 15

[127] J. De Leeuw and W. J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. In *Geometric Representations of Relational Data, Eds. J. C. Lingoes, E. Roskam, and I. Borg, Ann Arbor,MI: Mathesis Press,*, pages 735–752, 1977. 71

[128] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya. Learning unions of orthonormal bases with thresholded singular value decomposition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, page 293296, Philadelphia, PA, USA, 2005. 4

[129] K. Leung, M. van Stralen, A. Nemes, M. Voormolen, G. van Burken, M. Geleijnse, F. ten Cate, J. Reiber, N. de Jong, A. van der Steen, and J. Bosch. Sparse registration for three-dimensional stress echocardiography. *IEEE Transactions on Medical Imaging*, 27:1568–1579, 2008. 2

[130] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000. 4

[131] S. Z. Li. On discontinuity-adaptive smoothness priors in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:576–586, 1995. 71

[132] W. Li and J. Preisig. Estimation of rapidly time-varying sparse channels. *IEEE Journal of Oceanic Engineering*, 32:927–939, 2007. 2

[133] Y. Li, Shun-Ichi Amari, A. Cichocki, D. W. C. Ho, and S. Xie. Underdetermined blind source separation based on sparse representation. *IEEE Transactions on Signal Processing*, 54:423–437, 2006. 2

[134] Chung-Jr Lian, Ktian-Ftr Chen, Hong-Hui Chen, and Liang-Gee Chen. Lifting based discrete wavelet transform architecture for JPEG2000. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 445–448, Sydney, Australia, 2001. 2

[135] H. Linhart and W. Zucchini. *Model Selection*. Wiley, Oxford, England, 1986. 19, 50, 121

[136] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905. 7

[137] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, New York, 1973. 43

[138] F. Luisier, T. Blu, and M. Unser. A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on Image Processing*, 16:593–606, 2007. 21, 111

[139] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25:72–82, 2008. 2

[140] C. Lyzell, J. Roll, and L. Ljung. The use of nonnegative garrote for order selection of ARX models. In *Proc. IEEE Conference on Decision and Control*, pages 1974–1979, Cancun, Mexico, 2008. 88

[141] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:791–804, 2012. 4

[142] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17:53–69, 2008. 2

[143] A. Majumdar and R.K. Ward. Non-convex group sparsity: Application to color imaging. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 469–472, Dallas, Texas, U.S.A., 2010. 18, 41

[144] D. Malioutov, M. Cetin, and A. S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions on Signal Processing*, 53:3010–3022, 2005. 17, 39, 60

[145] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993. 13

[146] C. L. Mallows. Some comments on Cp. *Technometrics*, 15:661–675, 1973. 19

[147] L. Mancera and J. Portilla. L0-norm-based representation through alternate projections. In *Proc. IEEE International Conference on Image Processing*, pages 2089 – 2092, Atlanta, GA, U.S.A., 2006. 16

[148] O. Mangasarian. *Nonlinear Programming*. McGraw-Hill, New York, 1969. 43

[149] A. Marin, C. Chaux, J.-C. Pesquet, and P. Ciuciu. Image reconstruction from multiple sensors using stein's principle. application to parallel MRI. In *Proc. IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 465–468, Chicago, USA, 2011. 21, 111

[150] G. Marjanovic and V. Solo. l_q matrix completion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3885–3888, Kyoto, Japan, 2012. 15, 71

[151] D. Materassi, G. Innocenti, and L. Giarré. Reduced complexity models in the identification of dynamical networks: links with sparsification problems. In *Proc. IEEE Conference on Decision and Control*, pages 4796–4801, Shanghai, China, 2009. 18, 101, 102

[152] D. Materassi, M. V. Salapaka, and L. Giarrè. Relations between structure and estimators in networks of dynamical systems. In *Proc. IEEE Conference on Decision and Control*, pages 162–167, Orlando, FL, USA, 2011. 101, 102

[153] R. Mazumder, J. H. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106:1125–1138, 2011. 27, 34, 43

[154] R. H. Middleton and G. C. Goodwin. *Digital Control and Estimation: A Unified Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1990. 88

[155] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 2002. 12, 13

[156] K. Miller. Least squares methods for ill-posed problems with a prescribed bound. *SIAM Journal on Mathematical Analysis*, 1:52–74, 1970. 19

[157] M. Mishali and Y. C. Eldar. Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Transactions on Signal Processing*, 56:4692–4702, 2008. 17

[158] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten. Fast sparse representation based on smoothed l^0 norm. In *Proc. 7th International Conference on Independent Component Analysis and Signal Separation*, pages 389–396, London, UK, 2007. 16

[159] G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed l^0 norm. *IEEE Transactions on Signal Processing*, 57:289–301, 2009. 16, 72

[160] C. Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61:5678–5682, 2000. 101

[161] F. Mosteller and J. W. Tukey. *Data analysis and regression. A second course in statistics*. Addison-Wesley, 1968. 19

[162] S. Muthukrishnan. *Data Streams: Algorithms and Applications*. Now Publishers, Boston, 2005. 10

[163] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 26:917–922, 1977. 12

[164] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995. 10

[165] M. Naylora, L. Roseb, and B. Moyle. Topology of foreign exchange markets using hierarchical structure methods. *Physica A*, 382:199–208, 2007. 18, 101

[166] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Journal of Applied and Computational Harmonic Analysis*, 26:301–321, 2009. 13, 52, 53

[167] L. Ng and V. Solo. A data-driven method for choosing smoothing parameters in optical flow problems. In *Proc. IEEE International Conference on Image Processing*, pages 360–363, Washington, DC, USA, 1997. 21

[168] L. Ng and V. Solo. Selecting the neighbourhood size, shape, weights and model order in optical flow estimation. In *Proc. IEEE International Conference on Image Processing*, pages 600–603, Vancouver, BC, Canada, 2000. 21

[169] L. S. H. Ngia. Separable nonlinear least-squares methods for efficient offline and on-line modeling of systems using Kautz and Laguerre filters. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 48:562–579, 2001. 89

[170] B. Ninness and F. Gustafsson. A unifying construction of orthonormal bases for system identification. *IEEE Transactions Automatic Control*, 42:515–521, 1997. 89

[171] Y. Nurges. Laguerre models in problems of approximation and identification of discrete systems. *Automation and Remote Control*, 48:346–352, 1987. 89

[172] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. 4

[173] B. Ophir, M. Lustig, and M. Elad. Multi-scale dictionary learning using wavelets. *IEEE Journal of Selected Topics in Signal Processing*, 5:1014–1024, 2011. 4

[174] J. M. Ortega and W. C. Rheinboldt. Iterative solutions of nonlinear equations in several variables. *New York: Academic*, pages 253–255, 1970. 71, 75

[175] M. Osborne, B. Presnell, and B. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000. 14

[176] A.M. Ostrowski. *Solution of equations and systems of equations*. Academic Press, New York, 1960. 47, 79

[177] P. Heuberger P. Van den Hof and J. Bokor. Identification with generalized orthonormal basis functions - statistical analysis and error bounds. In *Proc. SYSZD94*, pages 207–212, Copenhagen, Denmark, 1994. 89

[178] J. Pakanen and S. Karjalainen. Estimating static heat flows in buildings for energy allocation systems. *Energy and Buildings*, 38:1044–1052, 2006. 101

[179] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig. Super-gaussian mixture source model for ICA. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 854–861, 2006. 73

[180] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao. Variational EM algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*, pages 1059–1066, 2006. 73

[181] A. Papachristodoulou and B. Recht. Determining interconnections in chemical reaction networks. In *Proc. American Control Conference*, pages 4872–4877, New York City, USA, 2007. 102

[182] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi. Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays. *IEEE Journal of Selected Topics in Signal Processing*, 2:275–285, 2008. 2

[183] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of American Statistical Association*, 104:735–746, 2009. 32, 36, 37

[184] W. B. Pennebaker and J. L. Mitchell. JPEG *Still Image Data Compression Standard*. Springer, 1993. 2

[185] M.D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M.E. Davies. Sparse representations in audio and music: From coding to source separation. *Proceedings of the IEEE*, 98:995–1005, 2010. 2

[186] L.C. Potter, E. Ertin, J.T. Parker, and M. Cetin. Sparsity and compressed sensing in radar imaging. *Proceedings of the IEEE*, 98:1006–1020, 2010. 2

[187] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado. Subset selection in noise based on diversity measure minimization. *IEEE Transactions on Signal Processing*, 51:760 – 770, 2003. 15, 21, 29, 50, 52

[188] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Transactions on Signal Processing*, 47:187 – 200, 1999. 8, 10, 15, 30

[189] M. Rasouli, D. Westwick, and W. Rosehart. Incorporating term selection into separable nonlinear least squares identification methods. In *Proc. Canadian Conference on Electrical and Computer Engineering*, pages 892–895, Vancouver, BC, 2007. 15, 88, 89, 97

[190] M. Rasouli, D. T. Westwick, and W. D. Rosehart. Incorporating term selection into nonlinear block structured system identification. In *Proc.*

American Control Conference, pages 3710–3715, Baltimore, MD, 2010. 15, 88, 89, 97

[191] S. Rickard and M. Fallon. The gini index of speech. In *Proc. of the 40th Annual Conference on Information Sciences and Systems*, Princeton, NJ, 2004. 8, 9, 11

[192] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, 2007. 19

[193] C. R. Rojas and H. Hjalmarsson. Sparse estimation based on a validation criterion. In *Proc. IEEE Conference on Decision and Control and European Control Conference*, pages 2825–2830, Orlando, FL, USA, 2011. 88

[194] R. Rubinstein, A.M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98:1045 –1057, 2010. 4

[195] R. Saab, R. Chartrand, , and Y. Özgür. Stable sparse approximations via nonconvex optimization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3885–3888, Las Vegas, Nevada, U.S.A., 2008. 16

[196] S. Saha, C.J. Long, E. Brown, E. Aminoff, M. Bar, and V. Solo. Hemodynamic transfer function estimation with Laguerre polynomials and confidence intervals construction, from functional magnetic resonance imaging (fMRI) data. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages III – 109–112, Quebec, Canada, 2004. 89

[197] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin. Compressive topology identification of interconnected dynamic systems via clustered orthogonal matching pursuit. In *Proc. IEEE Conference on Decision and Control*, pages 174–180, Orlando, FL, USA, 2011. 18, 101, 102

[198] B. M. Sanandaji, T. L. Vincent, M. B. Wakin, R. Tòth, and K. Poolla. Compressive system identification of LTI and LTV ARX models. In *Proc. IEEE Conference on Decision and Control and European Control Conference*, pages 791–798, Orlando, FL, USA, 2011. 88

[199] M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *Lecture Notes in Computer Science. Springer*, pages 286–297, 2007. 72

[200] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978. 19, 50, 121

[201] U. J. Schwarz. Mathematical-statistical description of the iterative beam removing technique (method CLEAN). *Astronomy and Astrophysics*, 65:345–356, 1978. 12, 13

[202] G. A. F. Seber. *Multivariate Observations*. J. Wiley, 1984. 17

[203] A. J. Seneviratne and V. Solo. On vector L0 penalized multivariate regression. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3613–3616, Kyoto, Japan, 2012. 27, 39, 97

[204] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997. 20

[205] X. Shen and H.-C. Huang. Optimal model assessment, selection, and combination. *Journal of the American Statistical Association*, 101:554–568, 2006. 20

[206] X. Shen, H.-C. Huang, and J. Ye. Adaptive model selection and assessment for exponential family models. *Technometrics*, 46:306–317, 2004. 20

[207] X. Shen and J. Ye. Adaptive model selection. *Journal of the American Statistical Association*, 97:210–221, 2002. 20

[208] M. Shi and V. Solo. Empirical choice of smoothing parameters in robust optical flow estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 349–352, Montreal, Quebec, Canada, 2004. 21, 111

[209] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981. 20

[210] C. D. Sigg, T. Dikk, and J. M. Buhmann. Speech enhancement using generative dictionary learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:1698–1712, 2012. 2

[211] T. O. Silva. On the determination of the optimal pole position of laguerre filters. *IEEE Transactions Signal Processing*, 43:2079–2087, 1995. 90

[212] D. B. Skalak. Prototype and feature selection by sampling and random mutation hill climbing algorithms. In *Proc. of the Eleventh International Conference on Machine Learning*, pages 293–301, 1994. 12

[213] V. Solo. A sure-fired way to choose smoothing parameters in ill-conditioned inverse problems. In *Proc. IEEE International Conference on Image Processing*, pages 89–92, Lausanne, Switzerland, 1996. 21

[214] V. Solo. Wavelet signal estimation in coloured noise with extension to transfer function estimation. In *Proc. IEEE Conference on Decision and Control*, pages 3940–3941, Tempa Florida, USA, 1998. 88, 89

[215] V. Solo. Selection of regularisation parameters for total variation denoising. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1653–1655, Phoenix, AZ, USA, 1999. 21, 111

[216] V. Solo. Automatic stopping criterion for anisotropic diffusion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3929–3932, Salt Lake City, Utah, USA, 2001. 21

[217] V. Solo. A fast automatic stopping criterion for anisotropic diffusion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages II–1661 – II–1664, Orlando, Florida, USA, 2002. 21, 111

[218] V. Solo. Signals in coloured noise: joint non-parametric estimation of signal and of noise spectrum. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 301–304, Hong Kong, China, 2003. 21, 111

[219] V. Solo. Selection of tuning parameters for support vector machines. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages V–237–V–240, Philadelphia, PA, USA, 2005. 21, 111

[220] V. Solo. A modified clean algorithm does l_1 denoising. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3665–3668, Las Vegas, Nevada, U.S.A., 2008. 13, 66

[221] V. Solo, C.J. Long, E.N. Brown, E. Aminoff, M. Bar, and S. Saha. fMRI signal modeling using laguerre polynomials. In *Proc. International Conference on Image Processing*, pages 2431–2434, Singapore, 2004. 89

[222] V. Solo and A. J. Seneviratne. The quadratic concave algorithm. *in preparation for submission to IEEE Transactions on Signal Processing*. 126

[223] V. Solo and M. Ulfarsson. Threshold selection for group sparsity. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3754–3757, Dallas, TX, USA, 2010. 112, 120

[224] R. V. Southwell. Stress-calculation in frameworks by the method of systematic relaxation of constraints: I and II. *Proceedings of the Royal Society of London. Series A*, 151:56–95, 1935. 13

[225] B. Sriperumbudur and G. Lanckriet. On the convergence of the concave-convex procedure. In *Advances in Neural Information Processing Systems*, pages 1759–1767, 2009. 77, 87

[226] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151, 1981. 19, 21, 111

[227] R. L. Stevenson, B. E. Schmitz, and E. J. Delp. Discontinuity preserving regularization of inverse visual problems. *IEEE Transactions on Systems, Man and Cybernetics*, 24:455–469, 1994. 71

[228] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B*, 36:111–147, 1974. 19

[229] G. Strang. The discrete cosine transform. *SIAM Review*, 41:135–147, 1999. 2

[230] G. Szego. *Orthogonal Polynomials*. American Mathematical Society Colloqimn Publication Vol. XXII, American Mathematical Society, Rhode Island, NY, 1939. 90

[231] K.S. Tang, K.F. Man, S. Kwong, and Q. He. Genetic algorithms and their applications. *IEEE Signal Processing Magazine*, 13:22–37, 1996. 12

[232] D. S. Taubman and M. W. Marcellin. JPEG *2000 : Image Compression Fundamentals, Standards, and Practice*. Springer, 2002. 2

[233] G. Temple. The general theory of relaxation methods applied to linear systems. *Proceedings of the Royal Society of London. Series A*, 169:476–500, 1938. 13

[234] N. Thakoor and J. Gao. Branch-and-bound for model selection and its computational complexity. *IEEE Transactions on Knowledge and Data Engineering*, 23:655–668, 2011. 12

[235] N. Thakoor, J. Gao, and V. Devarajan. Multibody structure-and-motion segmentation by branch-and-bound model selection. *IEEE Transactions on Image Processing*, 19:1393–1402, 2010. 12

[236] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996. 14, 28, 95, 106

[237] R. Tibshirani and K. Knight. The covariance inflation criterion for model selection. *Journal of the Royal Statistical Society: Series B*, 61:529–546, 1999. 20

[238] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001. 17

[239] I. Tošić and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28:27–38, 2011. 4

[240] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231 – 2242, 2004. 12

[241] J. A. Tropp. Algorithms for simultaneous sparse approximation. part II: Convex relaxation. *Signal Processing*, 86:589–602, 2006. 17, 39

[242] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006. 14

[243] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 51:1031–1051, 2006. 28

[244] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53:4655 – 4666, 2007. 12

[245] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. part I: Greedy pursuit. *Signal Processing*, 86:572–588, 2006. 17, 60

[246] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98:948–958, 2010. 12, 31

[247] P. Tseng. Convergence of block coordinate descent methods for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494, 2001. 43

[248] C. L. Tucker, J. F. Gera, and P. Uetz. Towards an understanding of complex protein networks. *Trends Cell Biol*, 11:102–106, 2001. 101

[249] B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005. 14

[250] M. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Transactions on Signal Processing*, 56:5804–5816, 2008. 21, 111

[251] M. O. Ulfarsson and V. Solo. Rank selection in noisy PCA with SURE and random matrix theory. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3317–3320, Las Vegas, Nevada, U.S.A., 2008. 21, 51

[252] M. O. Ulfarsson and V. Solo. Vector l_0 sparse variable PCA. *IEEE Transactions on Signal Processing*, 59:1949–1958, 2011. 43

[253] M.O. Ulfarsson and V. Solo. Sparse variable noisy PCA using l_0 penalty. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3950–3953, Dallas, Texas, U.S.A., 2010. 39

[254] D. Urban and T. Keitt. Landscape connectivity: A graph-theoretic perspective. *Ecology*, 82:1205–1218, 2001. 101

[255] H. Vafaie and K. De Jong. Robust feature selection algorithms, 1993. 12

[256] P. Valdes-Sosa, J. Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society*, pages 969–981, 2005. 102

[257] E. van den Berg and M. P. Friedlander. Theoretical and empirical results for recovery from multiple measurements. *IEEE Transactions on Information Theory*, 56:2516–2527, 2010. 18, 39

[258] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:19451959, 2005. 4

[259] C. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17:227–238, 1996. 72

[260] C. R. Vogel. Non-convergence of the l-curve regularization parameter selection method. *Inverse Problems*, 12:535–547, 1996. 19, 50

[261] A. Wagner and D. Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society London Series B*, 268:1803–1810, 2001. 18, 101

[262] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Yi Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:372–386, 2012. 2

[263] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14:651–667, 1977. 19

[264] B. Wahlberg. System identification using Laguerre models. *IEEE Transactions Automatic Control*, 36:551–562, 1991. 88, 89

[265] B. Wahlberg. System identification using Kautz models. *IEEE Transactions Automatic Control*, 39:1276–1282, 1994. 89

[266] B. Wahlberg and P. M. Makila. On approximation of stable linear dynamical systems using laguerre and kautz functions. *Automatica*, 32:693–708, 1996. 89

[267] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009. 14

[268] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature London*, 393:440–442, 1998. 101

[269] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003. 16, 72

[270] D. Wipf and S. Nagarajan. Iterative reweighted l_1 and l_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4:317–329, 2010. 72, 73

[271] D. Wipf and B. Rao. Sparse bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, 52:2153–2164, 2004. 17

[272] J. Wright, Yi Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98:1031–1044, 2010. 2

[273] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57:2479–2493, 2009. 71

[274] C.-J. Wu and D. W. Lin. A group matching pursuit algorithm for sparse channel estimation for OFDM transmission. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages IV 429–IV 432, Toulouse, France, 2006. 18

[275] T. Wu and K. Lange. The MM alternative to EM. *Statistical Science*, 25:492505, 2010. 75

[276] J. Yang, J. Wright, T. S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19:2861–2873, 2010. 2

[277] Y. Yang. Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92:937–950, 2005. 20

[278] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006. 14, 18, 32, 41, 52, 102, 106, 112, 115, 116

[279] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 4:915–936, 2003. 70

[280] M. M. Zavlanos, A. A. Julius, and S. P. Boyd. Identification of stable genetic networks using convex programming. In *Proc. American Control Conference*, pages 2755–2760, Seattle, Washington, USA, 2008. 102

[281] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. 14

[282] X. Zheng and W.-Y. Loh. Consistent variable selection in linear models. *Journal of the American Statistical Association*, 90:151–156, 1995. 20

[283] H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35:2173–2192, 2007. 112