SOLUTION OF LINEAR ILL-POSED PROBLEMS USING OVERCOMPLETE DICTIONARIES

by

PAWAN KUMAR GUPTA
M.S. University of Central Florida, 2015
M.Sc. Indian Institute of Technology, Dhanbad, 2010
B.Sc. University of Burdwan, 2008

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Mathematics
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2019

Major Professor: Marianna Pensky

ABSTRACT

In this dissertation, we consider an application of overcomplete dictionaries to the solution of general ill-posed linear inverse problems. In the context of regression problems, there has been an enormous amount of effort to recover an unknown function using such dictionaries. While some research on the subject has been already carried out, there are still many gaps to address. In particular, one of the most popular methods, lasso, and its versions, is based on minimizing the empirical likelihood and unfortunately, requires stringent assumptions on the dictionary, the so-called, compatibility conditions. Though compatibility conditions are hard to satisfy, it is well known that this can be accomplished by using random dictionaries. In the first part of the dissertation, we show how one can apply random dictionaries to the solution of ill-posed linear inverse problems with Gaussian noise. We put a theoretical foundation under the suggested methodology and study its performance via simulations and real-data example. In the second part of the dissertation, we investigate application of lasso to the linear ill-posed problems with non-Gaussian noise. We have developed theoretical background for application of lasso to such problems and studied its performance via simulations.

ACKNOWLEDGMENTS

I would like to thank my Ph.D. advisor, Dr. Marianna Pensky for her guidance and patience during my years of study at University of Central Florida (UCF). I also thank Dr. Hassan Foroosh, Dr. Jason Swanson, and, Dr. Teng Zhang , for taking the time to serve in my dissertation committee.

I would like to express my most profound appreciation to Dr. Ram Narayan Mohapatra for being a great mentor throughout my Ph.D.

I would also like to thank my close friends at UCF for their support since last few years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Linear Ill-posed Problems

In most of the field of science and technology, we always deal with the situation where we are not able to obtain our measurement(s) accurately; instead, we end up with inaccurate measurement(s) containing error. Unfortunately, obtaining error-free measurements in a typical situation is almost impossible, and we measure out entity of interest $\mathbf{f}$ as a noisy version of it $\mathbf{q}$.

Encountering inverse problems in real life is more frequent than it seems, especially in the modern era, with exponential growth in the field of science and technology. An example of such is calculating an image in X-ray computer tomography is an inverse problem. Inverse problems are some of the most important mathematical problems in science and mathematics as they tell us about the parameters which are not directly observable. Its applications spread through many fields, for example, signal processing, medical imaging, astronomy, machine learning, and many other areas.

Inverse problems are one of the classical problems in Linear Algebra of the form,

$$\mathbf{q} = \mathbf{Q}\mathbf{f}, \tag{1.1}$$

where we want to solve for $\mathbf{f}$ in $\mathbb{R}^n$ given $\mathbf{q}$ in $\mathbb{R}^m$. Here, $\mathbf{Q}$ is a $\mathbb{R}^{m\times n}$ matrix which is known as operator or observation matrix, and it depends on the specific device used for obtaining measurements. The problem (1.1) is called the inverse problem because we seek cause based on a result. The inverse problem is the direct negation of a forward problem, which is again of the form (1.1), but instead, we want to calculate $\mathbf{q}$ given $\mathbf{f}$. In forward problem, we seek results based on the cause.

In an inverse problem, the amplification of the error passes from the data to the obtained solution during calculation which actually is a critical feature of the inverse problem. In fact, in a genuine inverse problem, it is impossible to avoid some amplification of the error in computing a solution from the noisy version of the data. Now, computing a meaningful solution becomes even more difficult when the operator $\mathbf{Q}$ does not have an inverse, or some of its eigenvalues are very small. Under this condition, the inverse problem is called ill-posed. In case, the operator $\mathbf{Q}$ is linear, the same problem (1.1) is called linear ill-posed inverse problem.

1.2 General Problem under Consideration

In this dissertation, we address the general ill-posed linear inverse problem defined in (1.1) where $\mathbf{Q}$ is a bounded linear operator with an unbounded inverse, and the right-hand side of (1.1), $\mathbf{q}$ is measured with error. In particular we consider the equation,

$$\mathbf{y} = \mathbf{q} + \sigma\boldsymbol{\eta}, \quad \mathbf{q} = \mathbf{Q}\mathbf{f}, \tag{1.1}$$

where $\mathbf{y}, \mathbf{q}, \mathbf{f}, \boldsymbol{\eta} \in \mathbb{R}^n$, $\mathbf{Q} \in \mathbb{R}^{n\times n}$. Here, $\mathbf{y}$ is observed, $\mathbf{q}$ is unobserved, $\mathbf{f}$ is the function to be estimated, σ is the noise level, and $\boldsymbol{\eta}$ is the noise vector which follows some known distribution. The above problem is ill-posed since although the matrix $\mathbf{Q}$ in (1.1) is invertible, its lowest eigenvalue is very small, especially when the number of sample points n is relatively large. A general linear inverse problem can usually be reduced to matrix-form formulation by, either expanding $\mathbf{y}$ and $\mathbf{f}$ over some collection of basis functions or by measuring them at some set of points.

Solutions of statistical inverse problem (1.1) usually rely on reduction of the problem to the sequence model by carrying out the singular value decomposition (SVD) as in [8], [10], and [30] and references therein, or its relaxed version, the wavelet-vaguelette decomposition which was

proposed by Donoho in [13] and further studied by Abramovich and Silverman in [**?**]. Another general approach for the same inverse problem is Galerkin method with subsequent model selection as in [11]. The idea behind these methodologies is to represent the function of interest via an orthonormal basis, which is motivated by the operator $\mathbf{Q}$. The advantages of these methodologies are that these are asymptotically optimal in minimax sense. But these methods also have drawbacks as in many situations, the SVD decomposition of the linear operator is unknown, and hence these methods become inapplicable. Wavelet-vaguelette decomposition relies on relatively stringent conditions that are satisfied only for specific operators, mainly, of the convolution type. Also, wavelet-based methods have an advantage when one estimates a one-dimensional function defined on a finite interval. However, these methods do not perform that well in case either the function is defined on an infinite domain or the function is of several variables.

1.3 Notations

In this dissertation we use the following notations.

For any vector $\mathbf{t} \in \mathbb{R}^p$, denote its ℓ_2, ℓ_1, ℓ_0 and ℓ_∞ norms by, respectively, $\|\mathbf{t}\|$, $\|\mathbf{t}\|_1$, $\|\mathbf{t}\|_0$ and $\|\mathbf{t}\|_\infty$. For any matrix $\mathbf{A}$, denote its i^{th} row and j^{th} column by, $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$ respectively. Denote its spectral and Frobenius norms by, respectively, $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_2$.

Denote $\mathcal{P} = \{1, \cdots, p\}$. For any subset of indices $J \subseteq \mathcal{P}$, subset J^c is its complement in $\mathcal{P}$ and $|J|$ is its cardinality, so that $|\mathcal{P}| = p$. Let $\mathcal{L}_J = \text{Span}\left\{\boldsymbol{\varphi}_j,\ j \in J\right\}$. If $J \subset \mathcal{P}$ and $\mathbf{t} \in \mathbb{R}^p$, then $\mathbf{t}_J \in \mathbb{R}^{|J|}$ denotes reduction of vector $\mathbf{t}$ to subset of the indices J. Also, $\mathbf{\Phi}_J$ denotes the reduction of matrix $\mathbf{\Phi}$ to columns $\mathbf{\Phi}_{\cdot j}$ with $j \in J$.

Denote by $\lambda_{\min}(m; \mathbf{\Phi})$ and $\lambda_{\max}(m; \mathbf{\Phi})$ the minimum and the maximum restricted eigenvalues of

matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ given by

$$\lambda_{\min}(m;\mathbf{\Phi}) = \min_{\substack{\mathbf{t}\in\mathbb{R}^p \\ \|\mathbf{t}\|_0\leq m}} \frac{\mathbf{t}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{t}}{\|\mathbf{t}\|_2^2}, \quad \lambda_{\max}(m;\mathbf{\Phi}) = \max_{\substack{\mathbf{t}\in\mathbb{R}^p \\ \|\mathbf{t}\|_0\leq m}} \frac{\mathbf{t}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{t}}{\|\mathbf{t}\|_2^2}. \tag{1.1}$$

1.4 Organization of the Dissertation

The rest of the dissertation is organized as follows.

In Chapter 2, we provide a technical background of linear ill-posed problems along with methodologies used in past highlighting their drawbacks. We also discuss the random dictionaries and some special class of random variables along with centration inequalities. We also justify the application of random dictionaries in compressive sensing.

In Chapter 3, we discuss how one can apply lasso with a weighted penalty using overcomplete dictionaries for the solution of linear ill-posed inverse problems with Gaussian noise. In particular, we provide a theoretical justification of using random overcomplete dictionary for the solution of ill-posed problem with Gaussian noise. We also prove the efficiency of our approach over other methods like SVD in some settings by carrying out a limited simulation study followed by a real data example. This part of the research is complete and has been published in [20].

In Chapter 4, we consider the application of lasso with a weighted penalty for the solution of the linear ill-posed problem under non-Gaussian noise. In this case, we considered three non-Gaussian noise scenarios: Poisson, Binomial, and Chi-square. Here also we provide the theoretical justification of concentration inequalities and prove the efficiency of our approach over other methods like SVD in some settings by carrying out a limited simulation study.

In Chapter 5, we conclude our dissertation with discussion and future work.

CHAPTER 2: TECHNICAL BACKGROUND

This chapter provides a detailed discussion of background research done in the past for solving ill-posed linear inverse problem (1.1). We also discuss some of the very important concentration inequalities used in this chapter. These results had been proved and used in several articles and journals before.

2.1 Review of Linear Ill-posed Inverse Problems

In the past, a lot of effort has been made addressing linear inverse problems. In particular, if we consider the noise vector $\boldsymbol{\eta}$ in (1.1) with entries as standard Gaussian random variables, the inverse problem (1.1) is known as white noise model. The solution of this kind of statistical inverse problem usually relies on reducing it to the sequence model using singular value decomposition (SVD) or its relaxed version, the wavelet-vaguelette decomposition (WVD) proposed by Donoho in [13] and then it was further studied by Abramovich and Silverman in [**?**].

The advantage of the methodologies listed above is that they are asymptotically optimal in a minimax sense. The function of interest is usually represented via an orthonormal basis which is motivated by the form of matrix $\mathbf{Q}$. However, in spite of being minimax optimal in many contexts, these approaches have drawbacks. In particular, in practical applications, the number of observations n may be low while noise level σ high. In this situation, if the unknown vector $\mathbf{f}$ does not have a relatively compact and accurate representation in the chosen basis, the precision of the resulting estimator will be poor.

In the last decade, a great deal of effort was spent on the recovery of an unknown vector $\mathbf{f}$ in regression setting from its noisy observations using overcomplete dictionaries. In particular, if $\mathbf{f}$

has a sparse representation in some dictionary, a collection of vectors used for the representation of $\mathbf{f}$, then $\mathbf{f}$ can be recovered with a much better precision than, for example, when it is expanded over an orthonormal basis. The methodology is based on the idea that the error of an estimator of $\mathbf{f}$ is approximately proportional to the number of dictionary functions that are used for representing $\mathbf{f}$, therefore, expanding a function of interest over fewer dictionary elements decreases the estimation error. Similar advantages hold in the case of linear inverse problems (see [28]). However, in order to represent a variety of functions using a small number of dictionary elements, one needs to consider a dictionary of much larger size than the number of available observations, the, so-called, overcomplete dictionary.

In the past, a variety of appealing techniques have been developed for the solution of the above problems using overcomplete dictionaries such as likelihood penalization methods and greedy algorithms. Due to the computational convenience, lasso methodology and its variants gained popularity, and these have been used for the solution of a number of theoretical and applied statistical problems (see, e.g., [3], and also [5] and references therein). However, application of lasso is based on maximizing the likelihood and, unfortunately, relies on stringent assumptions on the dictionary $\{\boldsymbol{\varphi}_k\}_{k=1}^p$, the so-called, compatibility conditions, for a proof of its optimality. In regression set up, as long as compatibility conditions hold, lasso identifies a linear combination of the dictionary elements which represent the function of interest best of all at a "price" which is proportional to $\sigma\sqrt{n^{-1}\log p}$ where $\log$ stands for the natural logarithm and p is the dictionary size (see, e.g., [5]). Regrettably, while compatibility conditions may be satisfied for the vectors $\boldsymbol{\varphi}_j$ in the original dictionary, they usually do not hold for their images $\mathbf{Q}\boldsymbol{\varphi}_j$ due to contraction imposed by the operator $\mathbf{Q}$. Pensky [28] showed how lasso solution can be modified, so that it delivers an optimal solution, however, compatibility assumptions in [28] remain very complex and hard to verify.

In the recent years, it has been discovered that in the regression setting, one can satisfy compatibility conditions for lasso by simply using random dictionaries. In particular, Vershynin [33]

provided a variety of way for construction of such dictionaries, i.e., dictionaries comprised of random vectors. The purpose of [33], however, is that the methodology is intended for the recovery of a function which is directly observed. In Chapter 3 explain how random dictionaries can be adopted for the solution of ill-posed linear inverse problems.

2.2 Previous Methodologies

In this section we discuss the three most popular methodologies, Singular Value Decomposition, Galerkin's Method, and, Wavelet-vaguelette Decomposition which are rigorously used for the solution of inverse problems.

2.2.1 Singular Value Decomposition

For using SVD approach for solution of the inverse problem of the form (1.1), one needs knowledge of the SVD of the measurement matrix i.e. the operator $\mathbf{Q}$. Under this condition (but also in some other situations) the indirect observational model is derived to a model with direct observations and correlated data. Then the function of interest is estimated by minimizing the mean integrated squared risk (MISE) of the estimated function. The application of SVD approach is limited since the basis functions are motivated by the operator $\mathbf{Q}$ instead of the data or the function of interest. As a consequence, for example, if two completely different scientific field has same linear operator $\mathbf{Q}$, the same eigen basis functions will be used for function recovery although the function of interests could be quite different. It could cause the SVD estimator to perform poorly in broader perspective.

2.2.2 Galerkin's Method

Another common method for solving the same inverse problem is the Galerkin's method with subsequent model selection which is more appealing than SVD. In Galerkin's method one defines the approximation $\hat{\mathrm{f}}$ by solving a subsequent linear system. Moreover, if Q is a self-adjoint positive definite operator, the linear system becomes particularly simple to solve since the corresponding discretized operator is symmetric positive definite.

2.2.3 Wavelet-vaguelette Decomposition

The solution using wavelet-vaguelette decomposition proposed by Donoho rely on an orthogonal wavelet basis (ψ_i) and associated Riesz bases "vaguelettes" defined as

$$v_i = \beta_i \mathbf{Q}^{-1}\psi_i \text{ and } u_i = \beta_i (\mathbf{Q}^*)^{-1}\psi_i,$$

where the scaling co-efficient β_i depends on the order of ill-posedness of the operator $\mathbf{Q}$. Thus we obtain WVD as,

$$\mathbf{f} = \sum_{i=1}^{n} \beta_i^{-1} \langle \mathbf{Qf}, u_i \rangle \psi_i = \sum_{i=1}^{n} \beta_i^{-1} \langle \mathbf{Qf}, \psi_i \rangle v_i.$$

And, hence the WVD method leads to estimating the co-efficients in the above expansions from the observed data and applying a thresholding procedure.

The advantage of using all the three above methods are that they are asymptotically optimal in minimax sense and hence they gives best rates in the worst case scenario settings. The approach is to represent the function of interest as a linear combination of some orthonormal basis which is motivated by the type of the operator. But, inspite of its asymptotically optimal advantage the above

methods still faces two major drawbacks. First, the above methodology is difficult to implement because of the fact that in most of the cases the SVD of the linear operator is not known. On the other hand Wavelet-vaguelette decomposition depends on relatively stringent condition that are satisfied only for the operators of specific type, mainly of convolution type. Also, though wavelet-based methods perform well for functions of one variable in a finite interval, it doesn't perform same for a function of several variables [see, e.g., [26]]. The second drawback is that, the orthonormal dictionary used as representing the function, may not be rich enough and as result the chosen basis might not be able to represent the function of interest compact and accurately leading to poor estimation.

The above methods were deeply studied in past for solving ill-posed linear inverse problems under additive Gaussian noise which is also known as white noise. However, very little work has been done on the same problem for non-additive noise cases. For example, Poisson, Binomial and Chi-square noise (which is a specific case of Gamma noise) cases. Mathematical model for count and categorical data, such as Poisson arises in a variety of context, such as high-energy astrophysics or medical imaging, remote sensing data and only few to mention. It should also be noted that the above mentioned noise cases belong to the natural exponential families with a quadratic variance function (NEFQVF) as this class of functions has variance as maximum quadratic function of their mean. [4] suggested an unified treatment of these regression problems by using a mean-matching variance stabilizing transformation (VST) approach. The mean-matching VST transforms the relatively complicated problem of regression in exponential families into a standard homoscedastic Gaussian regression problem and then any good nonparametric Gaussian regression procedure can be used. For this purpose, they first grouped the data into many small sized bins, and then they applied the mean-matching VST to the binned data. [24] considered a similar problem for Poisson and multinomial noise case. They considered a multiscale factorization of a given data likelihood in analogy to the orthogonal wavelet decomposition.

Specifically in the Poisson noise framework, [17] proposed wavelet-based algorithm for estimating the deterministic discretized intensity function of an inhomogeneous one-dimensional Poisson process. The method was based on the asymptotic normality of a certain function of the Haar wavelet and scaling coefficients of the observed vector. The method first process the data using a nonlinear wavelet-based transformation, which is known as the Haar-Fisz transformation, and then treated the preprocessed vector as a signal with additive i.i.d. Gaussian noise with unit variance.

In the same framework, [34] proposed estimating the function of interest $\mathbf{f}$ from the data $\mathbf{y}$ using a regularized Poisson log-likelihood objective function. The idea was to maximize the log-likelihood while minimizing a penalty function that is proportional to the l_0-norm of the solution. They also presented risk bounds for recovery of a compressible signal from data.

[23] suggested a different approach by introducing a class of Bayesian multiscale models (BMSM's) for one dimensional inhomogeneous Poisson processes. These BMSM's were constructed using Recursive dyadic partitions (RDP's) within an entirely likelihood based frame work. Each RDP may be associated with a binary tree, and a new multi scale prior distribution was introduced for the unknown intensity through the placement of mixture distributions at each of the nodes of the tree. Then the concept of model mixing is applied to the complete collection of such trees. The advantage of this method was that not only it allows the inclusion of full location or scale information in the model, it induces both stationarity in the prior distribution and it enables a given intensity function to be approximated at the resolution of the data.

2.3 Random Matrices and Related Concentration Inequalities

Now we focus our discussion towards a special category of matrices known as random matrices. We address the reason for its suitability in so many applications in science and technology along

with in our study. We also discuss some very important concentration inequalities related to random matrices helpful throughout this dissertation.

2.3.1 Random Matrices

A random matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix with some or all of its entries as random variables following a probability distribution. The random matrices have wide applications in various branches of science and technology, out of which in this dissertation, we have only considered its application in Compressed sensing. It is very interesting that the random matrices are considered to be the best fit as a measurement matrix in compressed sensing by the researchers. This is due to the fact that the random matrices do satisfy the sufficient condition for a matrix to succeed for the purposes of compressed sensing and this special condition is known as the restricted isometry property (RIP). To satisfy this specific property a matrix needs all of its submatrices of given size be well-conditioned. This fits well in the circle of problems of the non-asymptotic random matrix theory. Now it is usual for one to wonder which specific types of random matrices satisfy the RIP. The answer is, all basic models of random matrices are nice restricted isometries. These include Gaussian and Bernoulli matrices, more generally all matrices with sub-Gaussian independent entries, and even more generally all matrices with sub-Gaussian independent rows or columns. Also, the class of restricted isometries includes random Fourier matrices, more generally random sub-matrices of bounded orthogonal matrices, and even more generally matrices whose rows are independent samples from an isotropic distribution with uniformly bounded coordinates.

As the dimensions m and n of the random matrix $\mathbf{A}$ grow to infinity, its spectrum tends to stabilize. This is manifested in several limit laws, which may be regarded as random matrix versions of the central limit theorem. Let us denote the singular values of $\mathbf{A}$ by $s_i(\mathbf{A})$ for $i = 1, 2, \cdots, n$. Note that the singular values are non-negative real numbers and they are often written in an non-

decreasing order as $s_1(\mathbf{A}) \geq \cdots \geq s_n(\mathbf{A}) \geq 0$. The largest singular value $s_1(\mathbf{A})$ is same as the operator norm of $\mathbf{A}$ which is defined as

$$\|A\|_{op} = \max\{\|\mathbf{A}\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\| = 1\}.$$

Many applications require estimates on the extreme singular values $s_{\max}(\mathbf{A}) := s_1(\mathbf{A}), s_{\min}(\mathbf{A}) := s_n(\mathbf{A})$. The smallest singular value is only of interest for tall matrices, since for wide matrices, $m < n$ and therefore, one automatically has $s_{\min}(\mathbf{A}) = 0$. An equivalent definition of $s_{\max}(\mathbf{A})$ and $s_{\min}(\mathbf{A})$ are respectively the smallest number C and the largest number c such that

$$c\|x\|_2 \leq \|Ax\|_2 \leq C\|x\|_2 \text{ for all } x \in R^n. \tag{2.1}$$

Geometrically, $\mathbf{A}$ can be interpreted as a linear operator from $\mathbb{R}^n \to \mathbb{R}^m$ and under its action, the Euclidean distance between any two points in $\mathbb{R}^n$ can get magnified by at most the factor $s_{\max}(\mathbf{A})$ or it can get shrinked by at most the factor $s_{\min}(\mathbf{A})$. In other words, the extreme singular values of $\mathbf{A}$ gives us the estimates of the upper and lower bounds of the distortion of the Euclidean geometry under the action of $\mathbf{A}$. If $s_{\max}(\mathbf{A}) \approx s_{\min}(\mathbf{A}) \approx 1$ then $\mathbf{A}$ acts as an approximate isometry, or more accurately an approximate isometric embedding of $l_2^n \to l_2^m$. Roughly speaking, in this case $\mathbf{A}$ does not effect the Euclidean geometry significantly and consequently $\mathbf{A}$ preserves the important properties of the Euclidean space. The extreme singular values can also be described in terms of the spectral norm of $\mathbf{A}$, which is by definition

$$\|\mathbf{A}\| = \|\mathbf{A}\|_{l_2^n \to l_2^m} = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|\mathbf{A}x\|_2}{\|x\|_2} = \sup_{x \in S^{n-1}} \|\mathbf{A}x\|_2.$$

(2.1) gives a link between the extreme singular values and the spectral norm: $s_{\max}(\mathbf{A}) = \|\mathbf{A}\|$ and $s_{\min}(\mathbf{A}) = \frac{1}{\|\mathbf{A}^+\|}$; if $\mathbf{A}$ is invertible then $\mathbf{A}^+ = \mathbf{A}^{-1}$.

Now, we are going to study some class of random variables which would be helpful in constructing suitable random matrices for our studies.

2.3.2 Sub-Gaussian Random Variables

Let us introduce a special class of random variables known as sub-Gaussian random variables which has very strong tail decay property. This class of variables are dominated by the Gaussian variables. In other words, the tail decay of the sub-Gaussian variables is at least as fast as the tail decay of the Gaussian variables.

Definition 2.3.1 (Sub-Gaussian random variables)**.** *A random variable X is called sub-Gaussian if it satisfies at least one of the following conditions:*

$$\mathbb{P}\{|X| \geq t\} \leq \exp(1 - \frac{t^2}{K_1^2}) \text{ for all } t \geq 0;$$

$$(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq K_2\sqrt{p}; \tag{2.2}$$

$$\mathbb{E}\exp\left(\frac{X^2}{K_3^2}\right) \leq e;$$

$$\mathbb{E}(tX) \leq \exp\left(t^2K_4^2\right) \text{ if } \mathbb{E}X = 0.$$

Note that, the above conditions are equivalent with parameters $K_i > 0, i = 1, \cdots, 4$ differing from each other by at most an absolute constant factor.

Definition 2.3.2 (Sub-Gaussian norm)**.** *The sub-Gaussian norm of a sub-Gaussian variable X is denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_2 in* (2.2)*.*

Equivalently, the sub-Gaussian norm can be defined as,

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-\frac{1}{2}} (\mathbb{E}|X|_p)^{\frac{1}{p}}.$$

Example 2.3.2.1. *Any standard normal random variable X is a sub-Gaussian random variable with the sub-Gaussian norm $\|X\|_{\psi_2} \leq C$ where C is an absolute constant. In general, if X is a centered normal random variable with variance σ^2, then X is sub-Gaussian with $\|X\|_{\psi_2} \leq C\sigma$. Also, a symmetric Bernoulli variable defined as, $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = \frac{1}{2}$ is a sub-Gaussian random variable with $\|X\|_{\psi_2} = 1$ since $|X| = 1$.*

More generally, if we consider any bounded random variable X, there exist some M such that $|X| \leq M$ almost surely which implies that X is a sub-Gaussian random variable with $\|X\|_{\psi_2} \leq M$. In other words, $\|X\|_{\psi_2} \leq \|X\|_{\infty}$.

Now, one of the quite remarkable property of the Gaussian variables is their rotation invariance. This means, given a finite number of independent centered Gaussian random variables, their sum is also a centered Gaussian random variable with variance as the individual sum of all the centered Gaussian random variables. Interestingly, sub-Gaussian random variables do have this property, although partially given by the following Lemma:

Lemma [33]: Let X_i for $i = 1, 2, \cdots, n$ be a finite number of independent centered sub-Gaussian random variables . Then $\sum_{i=1}^{n} X_i$ is also a centered sub-Gaussian random variable. Moreover,

$$\|\sum_{i=1}^{n} X_i\|_{\psi_2}^2 \leq C \sum_{i=1}^{n} \|X_i\|_{\psi_2}^2$$

where C is an absolute constant.

Using the above lemma, we can get the following inequality which provide us an upper bound for the large deviation of the finite sum of independent centered Gaussian random variables from its

mean.

Proposition [33]: Let $X_1, ..., X_m$ be independent centered sub-Gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1, ..., a_m) \in \mathbb{R}^m$ and every $t \geq 0$, we have

$$\mathbb{P}\{|\sum_{i=1}^{m} a_i X_i| \ \geq t\} \leq e \cdot \exp\{-\frac{ct^2}{K^2\|a\|_2^2}\}$$

where $c > 0$ is an absolute constant.

2.3.3 Sub-exponential Random Variables

The class of sub-Gaussian random variables is itself quite large but it leaves some of very useful random variable types which has heavier tail than Gaussian random variables. As an example, standard exponential random variables, a non-negative random variable with exponential tail decay as

$$\mathbb{P}(X \geq t) = \exp(-t), \ t \geq 0,$$

is one of them.

Definition 2.3.3 (Sub-exponential random variables)**.** *A random variable X is called sub-exponential if and only if it satisfies at least one of the following properties:*

$$\mathbb{P}\{|X| > t\} \leq \exp(1 - \frac{t}{K_1}) \ \textit{for all } t \geq 0.$$

$$(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq K_2 p \textit{ for all } p \geq 1.$$

$$\mathbb{E}\exp\left(\frac{X}{K_3}\right) \leq e.$$

Note that, the above properties are equivalent to each other with parameters $K_i > 0$ differing from each other by at most an absolute constant factor.

Definition 2.3.4 (Sub-exponential norm)**.** *The sub-exponential norm of X, denoted by* $\|X\|_{\psi_1}$*, is defined to be the smallest parameter* K_2*. In other words,*

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}}.$$

Example 2.3.4.1. *Chi-square distribution is an example of sub-exponential distribution.*

Sub-exponential random variables also hold large deviation inequalities similar to sub-Gaussian random variables.

Lemma [33]: Let $X_1, \cdots, X_m$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \cdots, a_m) \in \mathbb{R}^m$ and every $t \geq 0$, we have

$$\mathbb{P}\left(|\sum_{i=1}^{m} a_i X_i| \geq t\right) \leq 2 \cdot \exp\left[-c \min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $c > 0$ is an absolute constant.

Note that, the definitions of sub-Gaussian and sub-exponential random variables X do not require them to be centered. In order to use the lemma or theorem above we can always center X using the simple fact that if X is sub-Gaussian (or sub-exponential), then so is $X - \mathbb{E}X$. Also, in this case we have,

$$\|X - \mathbb{E}X\|_{\psi_1} \leq 2\|X\|_{\psi_1}, \text{ and } \|X - \mathbb{E}X\|_{\psi_2} \leq 2\|X\|_{\psi_2}.$$

Now we are carrying our discussion to higher dimensions. Therefore, we will discuss random vectors in $\mathbb{R}^n$ instead of random variable. A random vector is simply a vector with its entries as random variables following a probability distribution. While the concept of the mean $\mu = \mathbb{E}Z$ of

a random variable Z remains the same in higher dimensions, the second moment $\mathbb{E}Z^2$ is replaced by the $n \times n$ second moment matrix of a random vector $\mathbf{X}$, defined as,

$$\Sigma = \Sigma(\mathbf{X}) = \mathbb{E}\mathbf{X} \otimes \mathbf{X} = \mathbb{E}(\mathbf{X}\mathbf{X}^T),$$

where $\otimes$ denotes the outer product of vectors in $\mathbb{R}^n$. Similarly, the concept of variance $\text{Var}(Z) = \mathbb{E}(Z - \mu)^2 = \mathbb{E}Z^2 - \mu^2$ of a random variable is replaced in higher dimensions with the covariance matrix of a random vector $\mathbf{X}$, defined as,

$$\text{Cov}(\mathbf{X}) = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X}) \otimes (\mathbf{X} - \mathbb{E}\mathbf{X}) = \mathbb{E}\mathbf{X} \otimes \mathbf{X} - \mathbb{E}\mathbf{X} \otimes \mathbb{E}\mathbf{X}.$$

By translation, many questions can be reduced to the case of centered random vectors, for which $\mathbb{E}\mathbf{X} = 0$ and Cov $= \Sigma(X)$. We will also need a higher- dimensional version of unit variance:

2.3.4 Isotropic Random Vectors

Definition 2.3.5. ***(Isotropic random vectors)*** *A random vector $\mathbf{X} \in \mathbb{R}^n$ is called isotropic if $\mathbb{E}(\mathbf{X}\mathbf{X}^T) = \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of size n.*
Equivalently, we can define a isotropic random vector $\mathbf{X}$ to be a random vector in $\mathbb{R}^n$ so that,

$$\mathbb{E}\langle \mathbf{X}, \mathbf{y} \rangle^2 = \|\mathbf{y}\|^2 \text{ for all } \mathbf{y} \in \mathbb{R}^n.$$

For example, a vector in $\mathbb{R}^n$ with independent standard normal entries or independent symmetric Bernoulli entries are isotropic. More generally, consider a random vector $\mathbf{X}$ in $\mathbb{R}^n$ whose coordinates are independent random variables with zero mean and unit variance is an isotropic vector in $\mathbb{R}^n$. A more general version of the coordinate random vector is Frame.

Definition 2.3.6. ***(Frame)*** *A frame is a set of vectors $\{f_i\}_{i=1}^n$ in $\mathbb{R}^n$ which obeys an approximate*

Parsevals identity, i.e. there exist numbers $A, B > 0$ called frame bounds such that

$$A\|\mathbf{x}\|_2^2 \leq \sum_{i=1}^{n} |\langle f_i, \mathbf{x}\rangle|^2 \leq B\|\mathbf{x}\|_2^2$$

for all $\mathbf{x} \in \mathbb{R}^n$.

In particular if $A = B$, the set is called a tight frame. Thus, tight frames are generalizations of orthogonal bases without linear independence. Given a tight frame $\{u_i\}_{i=1}^n$ with bounds $A = B = n$, the random vector $\mathbf{X}$ uniformly distributed in the set $\{u_i\}_{i=1}^n$ is clearly isotropic in $\mathbb{R}^n$.

Suppose $\Sigma(\mathbf{X})$ is an invertible matrix, which means that the distribution of $\mathbf{X}$ is not essentially supported on any proper subspace of $\mathbb{R}^n$. Then $\Sigma(\mathbf{X})^{1/2}\mathbf{X}$ is an isotropic random vector in $\mathbb{R}^n$. Thus every non-degenerate random vector can be made isotropic by an appropriate linear transformation. This allows us to mostly focus on studying isotropic random vectors in the future.

Now, we are going to generalize the concepts of sub-Gaussian random variables to higher dimensions using one-dimensional marginals.

Definition 2.3.7. ***(Sub-Gaussian random vector)*** *A random vector $\mathbf{X}$ in $\mathbb{R}^n$ is sub-Gaussian if the one-dimensional marginals $\langle \mathbf{X}, \mathbf{x}\rangle$ are sub-Gaussian random variables for all $\mathbf{x}$ in $\mathbb{R}^n$.*

And in this case, the sub-Gaussian norm of the random vector $\mathbf{X}$ is defined as

$$\|\mathbf{X}\|_{\psi_2} = \sup_{\mathbf{x} \in S^{n-1}} \|\langle \mathbf{X}, \mathbf{x}\rangle\|_{\psi_2}.$$

The definitions of isotropic and sub-Gaussian distributions suggest that more generally, natural properties of high- dimensional distributions may be defined via one-dimensional marginals. This is a natural way to generalize properties of random variables to random vectors. For example, we shall call a random vector sub-exponential if all of its one-dimensional marginals are sub-

exponential random variables, etc.

2.3.5 Random Matrices with Independent Entries

Now we are going to study the extreme singular values of a random matrix. The most classical example of random matrices with independent entries are the Gaussian random matrices, matrices with entries as independent standard normal random variables. If, $\mathbf{A}$ is a $m \times n$ matrix with independent standard normal variables as entries, then the symmetric matrix $\mathbf{A}^*\mathbf{A}$ is called Wishart matrix; it is a higher-dimensional version of chi-square distributed random variables. Also, the simplest example of discrete random matrices is the Bernoulli random matrix $\mathbf{A}$ whose entries are independent symmetric Bernoulli random variables. In other words, Bernoulli random matrices are distributed uniformly in the set of all $m \times n$ matrices with ± 1 entries.

In this section we are going to discuss about the random matrices with independent and centered entries. Later we are going to discuss about more difficult cases when the rows and the coloumns of a random matrix are independent.

Now we have a theorem which provides us asymptotic version of the bounds of singular values of a random matrix as described above:

Theorem [33]: Let $\mathbf{A} = \mathbf{A}_{m,n}$ be an $m \times n$ random matrix whose entries are independent copies of a random variable with zero mean, unit variance, and finite fourth moment. Suppose that the dimensions m and n grow to infinity while the aspect ratio $\frac{n}{m}$ converges to a constant in $[0, 1]$. Then

$$s_{\min}(\mathbf{A}) = \sqrt{m} - \sqrt{n} + o(\sqrt{n}), s_{\max}(\mathbf{A}) = \sqrt{m} + \sqrt{n} + o(\sqrt{n}) \text{ almost surely.}$$

There is only one model of random matrices, Gaussian random matrices for which we have the exact non asymptotic version of the result which is given by the following theorem:

Theorem [33]: Let $\mathbf{A}$ be an $m \times n$ matrix whose entries are independent standard normal random variables. Then

$$\sqrt{m} - \sqrt{n} \leq \mathbb{E}(s_{\min}(\mathbf{A})) \leq \mathbb{E}(s_{\max}(\mathbf{A})) \leq \sqrt{m} + \sqrt{n}.$$

While the above Theorem is about the expectation of singular values, it also yields a large deviation inequality for them. It can be deduced formally by using the concentration of measure in the Gauss space.

Proposition [33]: Let f be a real valued Lipschitz function on $\mathbb{R}^n$ with Lipschitz constant K, i.e. $|f(x) - f(y)| \leq K\|x - y\|^2$ for all $x, y \in \mathbb{R}^n$ (such functions are also called K-Lipschitz). Let X be the standard normal random vector in $\mathbb{R}^n$. Then for every $t \geq 0$ one has $P\{f(X) - Ef(X) > t\} \leq exp(\frac{t^2}{2K^2})$.

Corollary [33]: Let $\mathbf{A}$ be an $m \times n$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(t^2/2)$ one has

$$\sqrt{m} - \sqrt{n} - t \leq \mathbb{E}\left(s_{\min}(\mathbf{A})\right) \leq \mathbb{E}\left(s_{\max}(\mathbf{A})\right) \leq \sqrt{m} + \sqrt{n} + t.$$

As we progress through the this dissertation, eventually we are going to see that it is more convenient to work with the $n \times n$ positive- definite symmetric matrix $\mathbf{A}^a st\mathbf{A}$ rather than with the original $m \times n$ matrix $\mathbf{A}$. Also, the normalized matrix $\bar{\mathbf{A}} = \frac{1}{\sqrt{m}}\mathbf{A}$ is an approximate isometry (which is our goal) if and only if $\bar{\mathbf{A}}^*\bar{\mathbf{A}}$ is an approximate identity (see [33]). In other words, if a matrix $\mathbf{B}$ that satisfies $\|\mathbf{B}^*\mathbf{B} - \mathbf{I}\| \leq \max(\delta, \delta^2)$ for some $\delta > 0$. Then $1 - \delta \leq s_{\min}(B) \leq s_{\max}(B) \leq 1 + \delta$. Conversely, if $\mathbf{B}$ satisfies for some $\delta > 0$ then $\|\mathbf{B}^*\mathbf{B} - I\| \leq 3\max(\delta, \delta^2)$.

2.3.6 General Random Matrices with Independent Entries

Now we pass to a more general model of random matrices whose entries are independent centered random variables with some general distribution (not necessarily normal). The largest singular value (the spectral norm) can be estimated by Latala's theorem for general random matrices with non-identically distributed entries: **Theorem [33]**: Let $\mathbf{A}$ be a random matrix whose entries a_{ij} are independent centered random variables with finite fourth moment. Then

$$\mathbb{E}\left(s_{\max}(\mathbf{A})\right) \leq C\left[\max_i\left(\sum_j \mathbb{E}a_{ij}^2\right)^{1/2} + \max_j\left(\sum_i \mathbb{E}a_{ij}^2\right)^{1/2} + \left(\sum_{i,j}\mathbb{E}a_{ij}^4\right)^{1/4}\right].$$

For almost square and square matrices, estimating the smallest singular value (known also as the hard edge of spectrum) is considerably more difficult. If $\mathbf{A}$ has independent entries, then indeed $s_{\min}(\mathbf{A}) \geq c(\sqrt{m} - \sqrt{n})$, and the following is an optimal probability bound:

Theorem [33]: Let $\mathbf{A}$ be an $m \times n$ random matrix whose entries are independent identically distributed sub-Gaussian random variables with zero mean and unit variance. Then for $\epsilon \geq 0$,

$$\mathbb{P}\left(s_{\min}(\mathbf{A}) \leq \epsilon(\sqrt{m} - \sqrt{n-1})\right) \leq (C\epsilon)^{m-n+1} + c^m,$$

where $C > 0$ and $c \in (0, 1)$ depend only on the sub-Gaussian norm of the entries.

This result gives an optimal bound for square matrices as well $(m = n)$.

Now, Vershynin gave theoritical proves for estimating extreme singular values for the random matrices with specific properties:

2.3.7 Random Matrices with Independent Sub-Gaussian Rows

Theorem [33]: Let $\mathbf{A}$ be an $m \times n$ matrix whose rows $\mathbf{A}_i$ are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^n$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(ct^2)$ one has

$$\sqrt{m} - C\sqrt{n} - t \leq s_{\min}(\mathbf{A}) \leq s_{\min}(\mathbf{A}) \leq \sqrt{m} + C\sqrt{n} + t.$$

Here $C = C_K, c = c_K > 0$ depend only on the sub-Gaussian norm $K = \max_i \|\mathbf{A}_i\|_{\psi_2}$ of the rows.

2.3.8 Random Matrices with Independent Heavy-tailed Rows

Theorem [33]: Let $\mathbf{A}$ be an $m \times n$ matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^n$. Let N be a number such that $\|A_i\|_2 \leq \sqrt{N}$ almost surely for all i. Then for every $t \geq 0$, one has

$$\sqrt{m} - t\sqrt{N} \leq s_{\min}(\mathbf{A}) \leq s_{max}(\mathbf{A}) \leq \sqrt{m} + t\sqrt{N}$$

with probability at least $1 - 2n\exp(ct^2)$, where $c > 0$ is an absolute constant.

2.3.9 Random Matrices with Independent Sub-Gaussian Columns

Theorem [33]: Let $\mathbf{A}$ be an $m \times n$ matrix $(m \geq n)$ whose columns A_i are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^m$ with $\|\mathbf{A}_j\|_2 = \sqrt{m}$ a. s. Then for every $t \geq 0$, the inequality holds

$$\sqrt{m} - C\sqrt{n} - t \leq s_{\min}(\mathbf{A}) \leq s_{\max}(\mathbf{A}) \leq \sqrt{m} + C\sqrt{n} + t$$

with probability at least $1 - 2\exp(ct^2)$, where $C = C_K, c = c_K > 0$ depend only on the sub-Gaussian norm $K = \max_j |\mathbf{A}_j\|_{\psi_2}$ of the columns.

The only significant difference between Theorems discussed above with independent rows and independent columns is that the latter requires normalization of columns, $\|\mathbf{A}_j\|_2 = \sqrt{m}$ almost surely.

2.3.10 Random Matrices with Independent Heavy-tailed Columns

Let $\mathbf{A}$ be an $m \times n$ matrix $(m \geq n)$ whose columns A_j are independent isotropic random vectors in $\mathbb{R}^m$ with $\|A_j\|_2 = \sqrt{m}$ a. s. Consider the incoherence parameter

$$N := \frac{1}{m}\mathbb{E}\max_{j\leq n}\sum_{k\in[n],k\neq j}\langle \mathbf{A}_j, \mathbf{A}_k\rangle^2.$$

Then $\mathbb{E}\|\frac{1}{m}\mathbf{A}^*\mathbf{A} - I\| \leq C_0\sqrt{\frac{N\log n}{m}}$. In particular,

$$\mathbb{E}\max_{j\leq n}|s_j(\mathbf{A}) - \sqrt{m}| \leq C\sqrt{N\log n}.$$

2.3.11 Restricted Isometry Property

In Compressive Sensing, when we use any measurement device it takes a signal $\mathbf{x} \in \mathbb{R}^n$ as input and returns the measurement as $y = \mathbf{A}\mathbf{x} \in \mathbb{R}^m$ as output where $\mathbf{A}$ is a $m \times n$ matrix acting as effect of the device on the original signal. In order to measure the data economically we seek m to be as small as possible, but also we should be able to recover the signal $\mathbf{x}$ from its inaccurate measurement $\mathbf{y}$. So, when we take very few measurements, i.e. when $m << n$, such matrices $\mathbf{A}$ are not one-to-one, so recovery of $\mathbf{x}$ from $\mathbf{y}$ is not possible for all signals $\mathbf{x}$. Practically, the

amount of information contained in the signal is often small. Mathematically this scenario is called sparsity of $\mathbf{x}$. In the simplest case, one assumes that $\mathbf{x}$ has few non-zero coordinates, say $|supp(\mathbf{x})| \leq k << n$. In this case, using any non-degenerate matrix $\mathbf{A}$ one can check that $\mathbf{x}$ can be recovered whenever $m > 2k$ using the optimization problem, $\min\{|supp(\mathbf{x})| : \mathbf{Ax} = \mathbf{y}\}$. This optimization problem is highly non-convex and generally NP-complete. So instead we a convex relaxation of this problem, $\min\{\|\mathbf{x}\|_1 : \mathbf{Ax} = \mathbf{y}\}$. A basic result in compressed sensing, due to Candes and Tao, is that for sparse signals $|supp(x)| \leq k$, the convex problem recovers the signal $\mathbf{x}$ from its measurement $\mathbf{y}$ exactly, provided that the measurement matrix $\mathbf{A}$ is quantitatively non-degenerate. Precisely, the non-degeneracy of $\mathbf{A}$ means that it satisfies the RIP with $\delta_{2k}(\mathbf{A}) \leq 0.1$.

Many signal classes have a low dimensional structure compared to the high-dimensional ambient space. Candes, Romberg, and Tao and Donoho have already shown the advantages of random projections for capturing information about sparse or compressible signals. As mentioned earlier a sufficient condition for a matrix to succeed for the purposes of compressed sensing is given by the RIP which demands that all submatrices of given size be well-conditioned.

Definition 2.3.8. ***(s-Sparse vectors)*** *A vector $\mathbf{x}$ in $\mathbb{R}^n$ is sparse if it has few non-zero coordinates i.e.*

$$|\textit{supp}(\mathbf{x})| << n.$$

In particular if

$$|\textit{supp}(\mathbf{x})| \leq s << n,$$

we call it s-sparse vector.

Definition 2.3.9. ***(Restricted Isometry Property (RIP))*** *An $m \times n$ matrix $\mathbf{A}$ satisfies the restricted isometry property of order $k \geq 1$ if there exists $\delta_k \geq 0$ such that the inequality*

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2$$

holds for all s-sparse $\mathbf{x}$ in $\mathbb{R}^n$.

The smallest number $\delta_k = \delta_k(\mathbf{A})$ is called the restricted isometry constant of $\mathbf{A}$. In other words, $\mathbf{A}$ has a RIP if $\mathbf{A}$ acts as an approximate isometry on all sparse vectors.

Clearly, $\delta_k(\mathbf{A}) = \max_{|\mathcal{T}|\leq k} \|\mathbf{A}^*_{\mathcal{T}}\mathbf{A}_{\mathcal{T}} - I_{\mathbb{R}^{\mathcal{T}}}\| = \max_{|\mathcal{T}|=\lfloor k \rfloor} \|\mathbf{A}^*_{\mathcal{T}}\mathbf{A}_{\mathcal{T}} - I_{\mathbb{R}^{\mathcal{T}}}\|$ where the maximum is over all subsets $\mathcal{T} \subset [n]$ with $|\mathcal{T}| \leq k$ or $|\mathcal{T}| = \lfloor k \rfloor$. The concept of restricted isometry can also be expressed via extreme singular values. $\mathbf{A}$ is a restricted isometry if and only if all $m \times k$ sub-matrices $\mathbf{A}_{\mathcal{T}}$ of $\mathbf{A}$ (obtained by selecting arbitrary k columns from $\mathbf{A}$) are approximate isometries. Indeed, for every $\delta \geq 0$, Versynin shows that the following two inequalities are equivalent up to an absolute constant: $\delta_k(\mathbf{A}) \leq \max(\delta, \delta^2)$;

$$1 - \delta \leq s_{\min}(\mathbf{A}_{\mathcal{T}}) \leq s_{\max}(\mathbf{A}_{\mathcal{T}}) \leq 1 + \delta$$

for all $|T| \leq k$.

Vershynin proved that $m \times n$ sub-Gaussian random matrices $\mathbf{A}$ are good restricted isometries: **Theorem [33]**: Let $\mathbf{A}$ be an $m \times n$ sub-Gaussian random matrix with independent rows or columns, which follows either of the two models above. Then the normalized matrix $\mathbf{B} = \frac{1}{\sqrt{m}}\mathbf{A}$ satisfies the following for every sparsity level $1 \leq k \leq n$ and every number $\delta \in (0, 1)$: if $m \geq C\delta^2 k \log(\frac{en}{k})$ then $\delta_k(\mathbf{B}) \leq \delta$ with probability at least $1 - 2\exp(c\delta^2 m)$. Here $C = C_K, c = c_K > 0$ depend only on the sub-Gaussian norm $K = \max_i \|\mathbf{A}_i\|_{\psi_2}$ of the rows or columns of $\mathbf{A}$.

The main reason that the above Theorem holds is that the random matrix $\mathbf{A}$ has a strong concentration property, i.e. that $\|\mathbf{Bx}\|_2 \approx \|\mathbf{x}\|_2$ with high probability for every fixed sparse vector $\mathbf{x}$. This concentration property alone implies the RIP, regardless of the specific random matrix model:

Proposition [33]: (Concentration implies restricted isometry). Let $\mathbf{A}$ be an $m \times n$ random matrix,

and let $k \geq 1, \delta \geq 0, \epsilon > 0$. Assume that for every fixed $\mathbf{x} \in \mathbb{R}^n$, $|supp(x)| \leq k$, the inequality

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2$$

holds with probability at least $1 - \exp(-\epsilon m)$. Then we have the following: if $m \geq C\epsilon^{-1}k\log(\frac{en}{k})$ then $\delta_k(\mathbf{B}) \leq 2\delta$ with probability at least $1 - exp(\frac{\epsilon m}{2})$. Here C is an absolute constant.

In words, the RIP can be checked on each individual vector $\mathbf{x}$ with high probability.

CHAPTER 3: SOLUTION OF LINEAR ILL-POSED PROBLEMS USING RANDOM DICTIONARIES

In this chapter, we address a specific case of our general ill-posed problem where the observations are corrupted by white noise i.e. the case where the noise vector $\boldsymbol{\eta}$ in (1.1) is a vector in $\mathbb{R}^n$ with i.i.d. standard normal entries.

3.1 Construction of the Lasso Estimator

Let $\boldsymbol{\Phi}$ the dictionary matrix with columns $\boldsymbol{\varphi}_j \in \mathbb{R}^n$, $j = 1, \cdots, p$, where p is possibly much larger than n and

$$\mathbf{f_t} = \sum_{j=1}^{p} t_j \boldsymbol{\varphi}_j = \boldsymbol{\Phi}\mathbf{t}. \tag{3.1}$$

Let $\boldsymbol{\theta}$ be the true vector of coefficients of expansion of $\mathbf{f}$ over the dictionary $\boldsymbol{\Phi}$ so that

$$\mathbf{f}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}. \tag{3.2}$$

Now we make the following assumption on our dictionary:

Assumption A0: Let vectors $\boldsymbol{\psi}_j$ be such that

$$\mathbf{Q}^T\boldsymbol{\psi}_j = \boldsymbol{\varphi}_j \text{ with} \|\boldsymbol{\psi}_j\|_\infty < \infty, \tag{3.3}$$

where $\mathbf{Q}^T$ is the transpose of matrix $\mathbf{Q}$, and $\boldsymbol{\Psi}$ be a matrix with columns $\boldsymbol{\psi}_j$, $j = 1, \cdots, p$. Then,

$$\mathbf{Q}^T\boldsymbol{\Psi} = \boldsymbol{\Phi} \quad \text{and} \quad \boldsymbol{\Psi} = \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\boldsymbol{\Phi}. \tag{3.4}$$

It should be noted that, although $\mathbf{f}$ is unknown,

$$\|\mathbf{f}-\mathbf{f_t}\|^2 = \|\mathbf{f}\|_2^2 + \mathbf{t}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{t} - 2\mathbf{t}^T\mathbf{\Phi}^T\mathbf{f} = \|\mathbf{f}\|_2^2 + \mathbf{t}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{t} - 2\mathbf{t}^T\mathbf{\Psi}^T\mathbf{Q}\mathbf{f} \tag{3.5}$$

is the sum of the three components where the first one is independent of $\mathbf{t}$, the second one is completely known, while the last term is of the form $2\mathbf{t}^T\mathbf{\Psi}^T\mathbf{Q}\mathbf{f} = 2\mathbf{t}^T\mathbf{\Psi}^T\mathbf{q}$ and, hence, can be estimated by $2\mathbf{t}^T\mathbf{\Psi}^T\mathbf{y}$. Let $\mathbf{z}$ be such that

$$\mathbf{\Psi}^T\mathbf{y} = \mathbf{\Phi}^T\mathbf{z}.$$

Therefore, expression $\mathbf{t}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{t} - 2\mathbf{t}^T\mathbf{\Psi}^T\mathbf{y}$ is minimized by the same vector $\mathbf{t}$ that minimizes $\|\mathbf{\Phi}\mathbf{t}-\mathbf{z}\|_2^2$ where

$$\mathbf{z} = (\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}\mathbf{\Psi}^T\mathbf{y}. \tag{3.6}$$

Let $\nu_j = \|\boldsymbol{\psi}_j\|_2$, $j = 1, \cdots, p$, and observe that ν_j is proportional to the standard deviation of the jth component of the vector $\mathbf{\Psi}^T\mathbf{y}$. Here ν_js can be viewed as a "cost" of using a dictionary element $\boldsymbol{\varphi}_j$ in the representation of $\mathbf{f}$. Considering

$$\mathbf{\Upsilon} = \text{diag}(\nu_1, \cdots, \nu_p) = \text{diag}(\|\boldsymbol{\psi}_1\|_2, \cdots, \|\boldsymbol{\psi}_p\|_2), \tag{3.7}$$

and following [28], we estimate the true vector of coefficients $\boldsymbol{\theta}$ as a solution of the quadratic optimization problem with the weighted lasso penalty

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \left\{ \|\mathbf{\Phi}\mathbf{t}-\mathbf{z}\|_2^2 + \alpha\|\mathbf{\Upsilon}\mathbf{t}\|_1 \right\}. \tag{3.8}$$

Here $\mathbf{z}$ is given by (3.6) and $\alpha \geq \alpha_0$ where

$$\alpha_0 = \sigma \sqrt{2\, n^{-1}\, (\tau + 1) \log p}. \tag{3.9}$$

Parameter $\tau > 0$ is related to the required probability bound (see formula (3.27) in Section 3.5 for details). Subsequently, we estimate the unknown solution $\mathbf{f}$ by $\widehat{\mathbf{f}} = \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}$.

Note that since we are interested in $\mathbf{f} = \mathbf{f}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$ rather than in the vector $\boldsymbol{\theta}$ of coefficients themselves, we are using lasso for the solution of the so-called prediction problem where it requires milder conditions on the dictionary. In fact, it is known (see [28]) that with no additional assumptions, for $\alpha \geq \alpha_0$, with probability at least $1 - 2p^{-\tau}$, one has

$$n^{-1} \left\| \mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f} \right\|_2^2 \leq \inf_{\mathbf{t}} \left[n^{-1} \left\| \mathbf{f}_{\mathbf{t}} - \mathbf{f} \right\|_2^2 + 4\alpha \sum_{j=1}^{p} \nu_j |t_j| \right]. \tag{3.10}$$

It is easy to see that if $\mathbf{t} = \boldsymbol{\theta}$, then $\mathbf{f}_{\mathbf{t}} = \mathbf{f}$. Then, with high probability, the error of the estimator $\mathbf{f}_{\widehat{\boldsymbol{\theta}}}$ is proportional to $\sigma \sqrt{n^{-1}\, (\tau + 1) \log p}\ \sum_j \nu_j$. This is the, so called, *slow lasso rate*. In order to attain the *fast lasso rate* proportional to $\sigma^2 n^{-1} \sum_j \nu_j^2$, one needs some kind of compatibility assumption.

3.2 Compatibility Condition

To obtain the fast lasso rate as discussed above, Pensky [28] formulated the following compatibility condition:

Assumption A1: Matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Upsilon}$ are such that for some $\mu > 1$ and any $J \subset \mathcal{P}$

$$\kappa^2(\mu, J) = \min \left\{ \mathbf{d} \in \mathcal{J}(\mu, J),\ \|\mathbf{d}\|_2 \neq 0 : \quad \frac{\mathbf{d}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} \mathbf{d} \cdot \mathrm{Tr}(\boldsymbol{\Upsilon}_J^2)}{\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1^2} \right\} > 0, \tag{3.11}$$

where $\mathcal{J}(\mu, J) = \left\{\mathbf{d} \in \mathbb{R}^p : \ \|(\boldsymbol{\Upsilon}\mathbf{d})_{J_*^c}\|_1 \leq \mu\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1\right\}$.

Pensky [28] proved that, under assumption (3.11), for $\alpha = \varpi\alpha_0$ where $\varpi \geq (\mu+1)/(\mu-1)$ and α_0 is defined in (3.9), with probability at least $1 - 2p^{-\tau}$, one has

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \leq \inf_{J \subseteq \mathcal{P}} \left[\|f - f_{\mathcal{L}_J}\|_2^2 + \frac{\sigma^2 K_0 (1+\varpi)^2(\tau+1)}{\kappa^2(\mu, J)} \frac{\log p}{n} \sum_{j \in J} \nu_j^2\right], \tag{3.12}$$

where $f_{\mathcal{L}_J} = \text{proj}_{\mathcal{L}_J} f$.

Note, however, that unless matrix $\boldsymbol{\Phi}$ has orthonormal columns, assumption (3.11) is hard not only to satisfy but even to verify since it requires checking it for every subset $J \in \mathcal{P}$. Indeed, sufficient conditions listed in Appendix A1 of [28] rely on the results of Bickel *et al.* [3] and require very stringent conditions on $\lambda_{\min}(m; \boldsymbol{\Phi})$ and entries $\boldsymbol{\Upsilon}$ in (4.9). In the next section, we suggest an alternative to this approach based on random dictionaries.

3.3 Lasso Solution To Linear Inverse Problems Using Random Dictionaries

An advantage of using random dictionary lies in the fact that one can ensure, with a high probability, that the dictionary satisfies a restricted isometry condition (see, e.g., [7] or [16]). In particular, if matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times p}$ satisfies the restricted isometry property of order $s \geq 1$, then $\lambda_{\min}(s; \boldsymbol{\Phi}) > 0$. The latter allows one to formulate the following results.

Theorem 3.3.1. *Let $\boldsymbol{\theta}$ be the solution of the optimization problem (3.8) with $\alpha \geq \alpha_0$ where α_0 is defined in (3.9). Let $\boldsymbol{\Phi} \in \mathbb{R}^{n \times p}$ be a random dictionary independent of $\mathbf{y}$ in (1.1). Denote*

$$J_* = \arg\min \left\{J \subset \mathcal{P} : \ n^{-1}\|\mathbf{f} - \mathbf{f}_{\mathcal{L}_J}\|_2^2 + K_0\alpha^2 \sum_{j \in J} \nu_j^2\right\}, \tag{3.13}$$

where $\mathbf{f}_{\mathcal{L}_J} = proj_{\mathcal{L}_J}\mathbf{f}$ and assume that $\boldsymbol{\Phi}$ is such that for some s, $1 \leq s \leq n/2$ and $\delta, \epsilon_1, \epsilon_2, \epsilon_3 \in$

$(0,1)$, *the following conditions hold*

$$\mathbb{P}\left(\lambda_{\min}(2s;\mathbf{\Phi}) \geq 1-\delta\right) \;\geq\; 1-\epsilon_1, \tag{3.14}$$

$$\mathbb{P}\left(|J_*| \leq s\right) \;\geq\; 1-\epsilon_2, \tag{3.15}$$

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\theta}}\|_0 \leq s\right) \;\geq\; 1-\epsilon_3, \tag{3.16}$$

If $K_0 \geq 4/(1-\delta)^2$ *in (3.13), then*

$$\mathbb{P}\left(\frac{1}{n}\left\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}}-\mathbf{f}\right\|_2^2 \leq \inf_{J\subseteq\mathcal{P}}\left[\frac{1}{n}\left\|\mathbf{f}-\mathbf{f}_{\mathcal{L}_J}\right\|_2^2 + K_0\alpha^2\sum_{j\in J}\nu_j^2\right]\right) \geq 1-2p^{-\tau}-\epsilon_1-\epsilon_2-\epsilon_3. \tag{3.17}$$

Note that for $\alpha=\alpha_0$ and $K_0 = 4/(1-\delta)^2$, under assumptions (3.14) – (3.16), (3.17) yields the following result

$$\mathbb{P}\left(\frac{1}{n}\left\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}}-\mathbf{f}\right\|_2^2 \leq \inf_{J\subseteq\mathcal{P}}\left\{\frac{1}{n}\left\|\mathbf{f}-\mathbf{f}_{\mathcal{L}_J}\right\|_2^2 + \frac{4\,\sigma^2}{n\,(1-\delta)^2}\sum_{j\in J}\nu_j^2\right\}\right) \geq 1-2p^{-\tau}-\epsilon_1-\epsilon_2-\epsilon_3. \tag{3.18}$$

As Lemma 3.3.2 below shows, assumption (3.14) can be guaranteed by choosing a dictionary of a particular type.

Lemma 3.3.2. *Let matrix* $\mathbf{\Phi}\in\mathbb{R}^{n\times p}$ *be independent of* $\mathbf{y}$ *and satisfy one of the following conditions:*

a) Matrix $\mathbf{\Phi}$ *has independent sub-Gaussian isotropic random rows;*

b) Matrix $\mathbf{\Phi}$ *has independent sub-Gaussian isotropic random columns with unit norms;*

c) Matrix $\mathbf{\Phi}$ *is obtained as* $\mathbf{\Phi}=(c\sqrt{n})^{-1}\mathbf{DW}$ *where* $\mathbf{W}\in\mathbb{R}^{m\times p}$ *is a matrix with i.i.d. standard Gaussian entries and columns of the matrix* $\mathbf{D}\in\mathbb{R}^{n\times m}$ *form a non-random* c*-tight frame so that for any vector* $\mathbf{x}$*, one has* $\mathbf{x}^T\mathbf{D}\mathbf{D}^T\mathbf{x}=c^2\|\mathbf{x}\|^2$.

If, for some $\delta \in (0,1)$ and $1 \leq s \leq n/2$, one has

$$n \geq C_1 \delta^{-2} s[\log(p/s) + 1], \tag{3.19}$$

then condition (3.14) holds with $\epsilon_1 \leq 2\exp(-C_2\delta^2 n)$. Here, C_1 and C_2 depend on the kind of sub-Gaussian variables that are involved in the formation of $\mathbf{\Phi}$ and are independent of n, m, p, s, and δ.

Finally, conditions (3.15) and (3.16) can be ensured by restricting the set of solutions $\mathbf{t}$ to vectors with cardinality at most s. In this case, $\epsilon_2 = \epsilon_3 = 0$ and the following corollary of Theorem 3.3.1 is valid.

Corollary 3.3.2.1. *Let $\boldsymbol{\theta}$ be the solution of optimization problem*

$$\widehat{\boldsymbol{\theta}} = \arg \min_{\mathbf{t}:\|\mathbf{t}\|_0 \leq s} \left\{ \|\mathbf{\Phi t} - \mathbf{z}\|_2^2 + \alpha \|\mathbf{\Upsilon t}\|_1 \right\}, \tag{3.20}$$

with $\alpha \geq \alpha_0$ where α_0 is defined in (3.9). Let $\mathbf{\Phi} \in \mathbb{R}^{n \times p}$ be one of the random dictionaries defined in Lemma 3.3.2. If, for some $\delta \in (0,1)$, condition (3.19) holds, then

$$\mathbb{P}\left(\frac{1}{n} \|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 \leq \inf_{\substack{J \subseteq \mathcal{P} \\ |J| \leq s}} \left[\frac{1}{n} \|\mathbf{f} - \mathbf{f}_{\mathcal{L}_J}\|_2^2 + \frac{4\alpha^2}{(1-\delta)^2} \sum_{j \in J} \nu_j^2 \right] \right) \geq 1 - 2p^{-\tau} - 2\exp(-C_2\delta^2 n), \tag{3.21}$$

where C_2 depends on the kind of sub-Gaussian variables that are involved in the formation of $\mathbf{\Phi}$ and is independent of n, m, p, s, and δ.

Note that case c) above offers a structured random dictionary since each of its elements is a linear combination of smooth functions.

3.4 Simulation Studies and Real-data Example

We substantiate the methodology discussed so far based on random dictionaries in the previous sections by carrying out a limited simulation study followed by a real-life example.

3.4.1 Simulation Setup

The experiment is being carried out by chosing three sample sizes: $n = 32$, $n = 64$ and $n = 128$ and two different signal to noise ratios (SNR): SNR = 3 and SNR = 5. We first generated the true vector $\mathbf{f}$ using `MakeSignal` program in the package Wavelab 850. We then generated the operator $\mathbf{Q}$ in (1.1) as $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where $\mathbf{U}$ is an $(n \times n)$ random orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with entries $\mathbf{\Lambda}_{ii} = 1/\sqrt{i}$, $i = 1, 2, \cdots, n$. $\mathbf{Q}$ was then used to obtain the unobserved vector $\mathbf{q}$ as

$$\mathbf{q} = \mathbf{Q}\mathbf{f}.$$

Finally, by adding Gaussian noise to $\mathbf{q}$ we generated our data $\mathbf{y}$. in order to do that, we first chose particular values of the Signal to Noise Ratio (SNR) and obtained σ as the ratio of the standard deviation of $\mathbf{q}$ and the SNR. Vector $\mathbf{y}$ was then calculated at n observation points as $\mathbf{y} = \mathbf{q} + \sigma\,\boldsymbol{\eta}$ where $\boldsymbol{\eta} \in \mathbb{R}^n$ is a standard normal vector. Our simulation is carried out for two noise levels: SNR = 3, relatively high noise and SNR = 5, relatively low noise.

3.4.2 Implementation Details

We compared the performance of the estimators of $\mathbf{f}$ based on random dictionaries with the estimator of $\mathbf{f}$ based on the SVD. For our simulations we have created three different $n \times p$ random dictionaries with $p = 5000$: (a) purely random dictionaries with, respectively, the i.i.d. standard

Gaussian entries and the i.i.d. sparse Bernoulli entries; (b) the fusion of the fixed dictionary and the random dictionary that follows case c) in Lemma 3.3.2 with $\mathbf{D}$ being the Haar dictionary. The sparse Bernoulli variable is defined as

$$\mathbf{X} = \begin{cases} -\sqrt{\frac{3}{n}} & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ \sqrt{\frac{3}{n}} & \text{with probability } \frac{1}{6}. \end{cases} \tag{3.22}$$

For creating the fusion dictionary, we first generated the orthogonal matrix of Haar wavelet transform $\mathbf{D}$ using `MakeWavelet` function, so that $m = n$ and $c = 1$. Then we obtained the dictionary $\mathbf{\Phi}$ following part c) of the Lemma 3.3.2 using the $n \times p$ matrix $\mathbf{W}$ with the i.i.d. normal entries.

We then obtained matrix $\mathbf{\Psi}$ of the inverse images as the numerical solution of the exact equation $\mathbf{Q}^T\mathbf{\Psi} = \mathbf{\Phi}$ and calculated vector $\mathbf{z}$ with elements (3.6). For the sake of obtaining a solution of the optimization problem (3.8), we used function `LassoWeighted` in SPAMS MatLab toolbox (see [27]).

In order to evaluate the value of the lasso parameter α, we first calculated $\alpha_{\max}$ as the value of α that guarantees that all coefficients in the model vanish. We then created a grid of the values of α as $\alpha_k = \alpha_{\max} * k/N$, $k = 1, \cdots, N$, with $N = 200$. Using this grid values of α, we obtained a collection of estimators $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\alpha_k)$. Now we needed to find most appropriate value of the tuning parameter α or in other words appropriate k which provides us the most accurate estimation. In order to do that, we estimated α as $\hat{\alpha} = \alpha_{\widehat{k}}$ in two ways: one using the oracle value of α and another using the estimated value of α. We found oracle value of α as $\alpha_{oracle} = \alpha_{\max} * \widehat{k}_{oracle}/N$ using the value $\widehat{k}_{oracle}$ that guarantees the most accurate estimator of $\mathbf{f}$:

$$\widehat{k}_{oracle} = \arg\min_k \|\mathbf{f} - \mathbf{\Phi}\widehat{\boldsymbol{\theta}}(\alpha_k)\|_2.$$

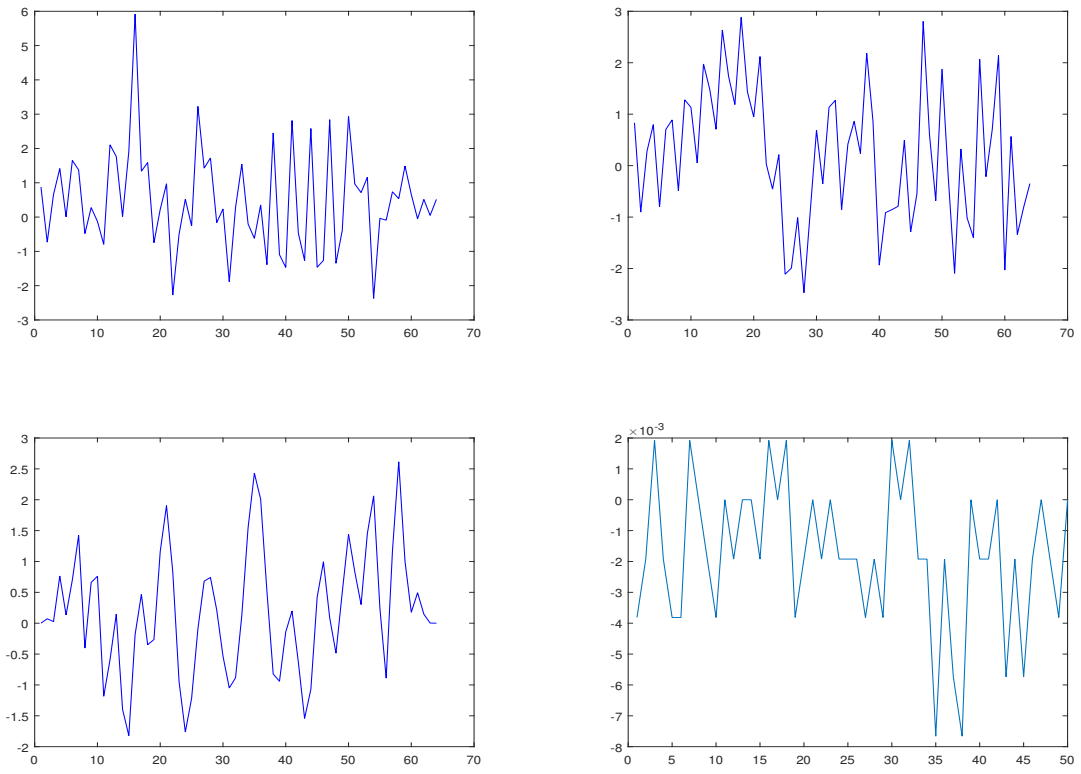

Figure 3.1: Test signals: `WernerSorrows` (top left), `MishMash` (top right), `Chirps` (bottom left) with $n = 64$ and `Bird's twitter` (bottom right) with $n = 50$.

Since the vector $\mathbf{f}$ is unavailable in real-life, we find the estimated value $\hat{\alpha}_{est} = \alpha_{\max} * \widehat{k}_{est}/N$ of α using

$$\widehat{k}_{est} = \arg\min_k \frac{1}{n}\|\mathbf{y} - \hat{\mathbf{q}}(\alpha_k)\|_2^2 + 2\sigma^2 n^{-1}\hat{p}_k,$$

where $\hat{\mathbf{q}}(\alpha_k) = \mathbf{Q}\mathbf{\Phi}\widehat{\boldsymbol{\theta}}(\alpha_k)$ is the estimator of $\mathbf{q}$ based on the lasso estimator obtained with the parameter α_k and $\hat{p}_k$ is the number of nonzero components of $\widehat{\boldsymbol{\theta}}(\alpha_k)$.

We compared the precision of the estimators $\hat{\mathbf{f}}_{RN}$, $\hat{\mathbf{f}}_{RB}$, $\hat{\mathbf{f}}_{RH}$ of $\mathbf{f}$ based, respectively, on Gaussian, Bernoulli and Haar fusion random dictionaries described above with $\hat{\mathbf{f}}^{SVD}_{oracle}$, the estimator based on the singular value decomposition (SVD). Precision of an estimator $\hat{\mathbf{f}}$ is measured by the relative

error

$$R(\hat{\mathbf{f}}) = \|\hat{\mathbf{f}} - \mathbf{f}\|/\|\mathbf{f}\|, \tag{3.23}$$

averaged over 50 simulation runs (with the standard deviations listed in parentheses). Initially, we considered wavelet estimator of $\mathbf{f}$ using Daubechies wavelet of order 8, but we discarded it due to its poor performance with respect to the estimators considered for comparison. For finding $\hat{\mathbf{f}}^{SVD}_{oracle}$, we used the oracle number K_{oracle} of eigenbasis functions. We obtained K_{oracle} as the number of eigenbasis functions that minimizes the difference between $\hat{\mathbf{f}}^{SVD}_{oracle}$ and the true function $\mathbf{f}$ which is unavailable in a real-life setting.

3.4.3 Simulation Results

We documented the accuracies of all the estimators used for comparison in Table 3.1 below. For all the three estimators based on random dictionaries, we report the errors with both the oracle and the estimated values of α, $\hat{\mathbf{f}}^{lasso}_{RN,oracle}$, $\hat{\mathbf{f}}^{lasso}_{RB,oracle}$, $\hat{\mathbf{f}}^{lasso}_{RH,oracle}$ and $\hat{\mathbf{f}}^{lasso}_{RN,cv}$, $\hat{\mathbf{f}}^{lasso}_{RB,cv}$, $\hat{\mathbf{f}}^{lasso}_{RH,cv}$, respectively. In our experiment we used three types of test signals `WernerSorrows`, `MishMash` and `Chirps` as shown in Figure 3.1.

Table 3.1 shows that the random dictionary based estimators have higher accuracy in estimating the original signals than the SVD estimator, they showed approximately 5-10% improvement over average errors than the oracle SVD estimator. The advantage of using random dictionaries is more noticeable when n is small ($n = 32$) and the noise level is high ($SNR = 3$). In addition, the improvement of $\hat{\mathbf{f}}^{lasso}_{RN,oracle}$, $\hat{\mathbf{f}}^{lasso}_{RB,oracle}$ and $\hat{\mathbf{f}}^{lasso}_{RH,oracle}$ over $\hat{\mathbf{f}}^{SVD}_{oracle}$ is more significant than that of $\hat{\mathbf{f}}^{lasso}_{RN,cv}$, $\hat{\mathbf{f}}^{lasso}_{RB,cv}$ and $\hat{\mathbf{f}}^{lasso}_{RH,cv}$ since the latter estimators loose accuracy because of suboptimal choices of the parameter α. Nevertheless, in the majority of cases, they still exhibit better precision than $\hat{\mathbf{f}}^{SVD}_{oracle}$ although this is not entirely fair comparison since $\hat{\mathbf{f}}^{SVD}_{oracle}$ is based on the oracle choice of parameter K. This is due to the fact that large random dictionaries provide a more sparse representation of $\mathbf{f}$.

Table 3.1: The average values of the errors $R(\hat{\mathbf{f}})$ evaluated over 50 simulation runs of the estimators for various test signals (standard deviations of the errors are listed in the parentheses).

$WernerSorrows$

	$\hat{\mathbf{f}}_{lasso}^{RN,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RN,cv}$	$\hat{\mathbf{f}}_{lasso}^{RB,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RB,cv}$	$\hat{\mathbf{f}}_{lasso}^{RH,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RH,cv}$	$\hat{\mathbf{f}}_{oracle}^{SVD}$
$n = 32$,	0.3416	0.3567	0.3416	0.3566	0.3365	0.3466	0.3587
$SNR = 3$	(0.0497)	(0.0534)	(0.0512)	(0.0559)	(0.0504)	(0.0498)	(0.0502)
$n = 32$,	0.2670	0.2752	0.2651	0.2759	0.2645	0.2725	0.2797
$SNR = 5$	(0.0447)	(0.0440)	(0.0455)	(0.0443)	(0.0433)	(0.0439)	(0.0440)
$n = 64$,	0.3778	0.3944	0.3814	0.3969	0.3739	0.3850	0.3970
$SNR = 3$	(0.0332)	(0.0351)	(0.0308)	(0.0349)	(0.0299)	(0.0331)	(0.0375)
$n = 64$,	0.2047	0.2083	0.2056	0.2084	0.2052	0.2077	0.2113
$SNR = 5$	(0.0228)	(0.0227)	(0.0234)	(0.0236)	(0.0227)	(0.0233)	(0.0233)
$n = 128$,	0.4066	0.4285	0.4079	0.4292	0.4054	0.4164	0.4388
$SNR = 3$	(0.0287)	(0.0304)	(0.0278)	(0.0312)	(0.0277)	(0.0294)	(0.0311)
$n = 128$,	0.2632	0.2687	0.2623	0.2682	0.2623	0.2651	0.2717
$SNR = 5$	(0.0177)	(0.0188)	(0.0172)	(0.0189)	(0.0176)	(0.0183)	(0.0191)

$MishMash$

	$\hat{\mathbf{f}}_{lasso}^{RN,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RN,cv}$	$\hat{\mathbf{f}}_{lasso}^{RB,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RB,cv}$	$\hat{\mathbf{f}}_{lasso}^{RH,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RH,cv}$	$\hat{\mathbf{f}}_{oracle}^{SVD}$
$n = 32$,	0.3934	0.4093	0.3951	0.4096	0.3932	0.4045	0.4347
$SNR = 3$	(0.0590)	(0.0634)	(0.0582)	(0.0621)	(0.0617)	(0.0615)	(0.0636)
$n = 32$,	0.2562	0.2615	0.2573	0.2628	0.2569	0.2613	0.2660
$SNR = 5$	(0.0428)	(0.0420)	(0.0408)	(0.0412)	(0.0403)	(0.0407)	(0.0399)
$n = 64$,	0.4621	0.4824	0.4642	0.4817	0.4632	0.4788	0.4974
$SNR = 3$	(0.0392)	(0.0465)	(0.0416)	(0.0448)	(0.0408)	(0.0457)	(0.0468)
$n = 64$,	0.2291	0.2326	0.2291	0.2324	0.2279	0.2312	0.2342
$SNR = 5$	(0.0225)	(0.0237)	(0.0213)	(0.0226)	(0.0221)	(0.0222)	(0.0224)
$n = 128$,	0.4042	0.4152	0.4037	0.4147	0.4040	0.4099	0.4277
$SNR = 3$	(0.0311)	(0.0323)	(0.0301)	(0.0326)	(0.0311)	(0.0318)	(0.0328)
$n = 128$,	0.2676	0.2712	0.2672	0.2711	0.2675	0.2696	0.2735
$SNR = 5$	(0.0193)	(0.0190)	(0.0191)	(0.0191)	(0.0187)	(0.0186)	(0.0194)

$Chirps$

	$\hat{\mathbf{f}}_{lasso}^{RN,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RN,cv}$	$\hat{\mathbf{f}}_{lasso}^{RB,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RB,cv}$	$\hat{\mathbf{f}}_{lasso}^{RH,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RH,cv}$	$\hat{\mathbf{f}}_{oracle}^{SVD}$
$n = 32$,	0.3497	0.3603	0.3513	0.3603	0.3502	0.3613	0.3746
$SNR = 3$	(0.0512)	(0.0525)	(0.0490)	(0.0486)	(0.0498)	(0.0507)	(0.0494)
$n = 32$,	0.2336	0.2421	0.2342	0.2407	0.2335	0.2400	0.2454
$SNR = 5$	(0.0375)	(0.0373)	(0.0364)	(0.0364)	(0.0372)	(0.0361)	(0.0360)
$n = 64$,	0.3932	0.4068	0.3935	0.4072	0.3938	0.4017	0.4246
$SNR = 3$	0.0365)	(0.0399)	(0.0388)	(0.0409)	(0.0401)	(0.0399)	(0.0386)
$n = 64$,	0.2702	0.2749	0.2700	0.2745	0.2691	0.2734	0.2756
$SNR = 5$	(0.0310)	(0.0341)	(0.0323)	(0.0347)	(0.0316)	(0.0337)	(0.0343)
$n = 128$,	0.4065	0.4146	0.4054	0.4144	0.4055	0.4144	0.4266
$SNR = 3$	(0.0302)	(0.0325)	(0.0296)	(0.0303)	(0.0303)	(0.0299)	(0.0307)
$n = 128$,	0.2545	0.2579	0.2548	0.2575	0.2545	0.2579	0.2593
$SNR = 5$	(0.0191)	(0.0188)	(0.0189)	(0.0188)	(0.0189)	(0.0190)	(0.0181)

Table 3.2: The average values of the relative errors $R(\hat{\mathbf{f}})$ evaluated over 50 simulation runs of the estimators for the test signal (standard deviations of the relative errors are listed in the parentheses).

Bird's twitter

	$\hat{\mathbf{f}}_{lasso}^{RB1,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RB1,cv}$	$\hat{\mathbf{f}}_{lasso}^{RB2,oracle}$	$\hat{\mathbf{f}}_{lasso}^{RB2,cv}$	$\hat{\mathbf{f}}_{oracle}^{SVD}$
$n = 50$, $SNR = 3$	0.4266 (0.0557)	0.4458 (0.0618)	0.4218 (0.0551)	0.4394 (0.0598)	0.4319 (0.0307)
$n = 50$, $SNR = 5$	0.2711 (0.0354)	0.2772 (0.0365)	0.2692 (0.0365)	0.2758 (0.0363)	0.2795 (0.0341)
$n = 50$, $SNR = 7$	0.1972 (0.0261)	0.2010 (0.0266)	0.1960 (0.0269)	0.1986 (0.0269)	0.2014 (0.0260)

To study the performance of the suggested method in a practical setting, we used a real-life signal for $\mathbf{f}$ that consists of a bird's twitter available as an audio signal on the internet at `http://www.externalharddrive.com/waves/animal/index.html`. We sampled from the signal to generate our true signal with lengths $n = 50$ and applied an averaging matrix operator $\mathbf{Q}$ to obtain $\mathbf{q}$. Here we chose the operator $\mathbf{Q}$ as the Toeplitz matrix with unit diagonal entries, upper diagonal entries equal to 0.5 and the rest of the entries equal to zero. We then finally obtained the noisy observations $\mathbf{y}$ by adding Gaussian random noise to $\mathbf{q}$ as before. For our study we used $p = 3000$ and considered three noise levels: $SNR = 3$ (high noise level), $SNR = 5$ (moderate noise level), and $SNR = 7$ (low noise level). In this case, we used estimators of $\mathbf{f}$ based on two random dictionaries, the sparse Bernoulli dictionary defined in (3.22) and the symmetric Bernoulli dictionary with i.i.d. random entries given by

$$\mathbf{X} = \begin{cases} 1/n & \text{with probability } 1/2 \\ -1/n & \text{with probability } 1/2. \end{cases} \tag{3.24}$$

We denote estimators based on dictionaries (3.22) and (3.24) by $\hat{\mathbf{f}}_{RB1}$ and $\hat{\mathbf{f}}_{RB2}$, respectively, and characterize estimation precision by $R(\hat{\mathbf{f}})$.

Table 3.2 below compares the accuracies of the random dictionary based estimators with the SVD

estimator using the relative error (3.23) averaged over 50 simulation runs (with the standard deviations of the precision listed in parentheses) for the real signal bird's twitter. Here also, for all the estimators based on random dictionaries, we report the errors of the estimators with both the oracle and the estimated values of α, denoted by $\hat{\mathbf{f}}^{lasso}_{RB1,oracle}$, $\hat{\mathbf{f}}^{lasso}_{RB2,oracle}$ and $\hat{\mathbf{f}}^{lasso}_{RB1,cv}$, $\hat{\mathbf{f}}^{lasso}_{RB2,cv}$, respectively. Similar to our simulation studies based on artificial test signals, from Table 3.2 it follows that all the random dictionary based estimators have higher accuracy than the SVD estimators for the real-life signal with approximately 5% better accuracy and the random dictionary estimators performed better for low sample size (n = 50) and high noise level (SNR = 3). Also, similarly to the case of artificial signals, the advantage of $\hat{\mathbf{f}}^{lasso}_{RB1,oracle}$ and $\hat{\mathbf{f}}^{lasso}_{RB2,oracle}$ over $\hat{\mathbf{f}}^{SVD}_{oracle}$ is more significant than that of $\hat{\mathbf{f}}^{lasso}_{RB1,cv}$ and $\hat{\mathbf{f}}^{lasso}_{RB2,cv}$ since the latter estimators loose accuracy because of suboptimal choices of the parameter α. Nevertheless, in the majority of cases, $\hat{\mathbf{f}}^{lasso}_{RB1,cv}$ and $\hat{\mathbf{f}}^{lasso}_{RB2,cv}$ still exhibit better precisions than $\hat{\mathbf{f}}^{SVD}_{oracle}$ although this is not entirely fair comparison since $\hat{\mathbf{f}}^{SVD}_{oracle}$ is based on the oracle choice of parameter K.

3.5 Proofs

Proof of Theorem 3.3.1. The beginning of the proof is similar to the proof of Lemma 2 in [28]. However, for completeness, we provide the complete proof here.

Let $\boldsymbol{\theta}$ be the true parameter vector so that $\mathbf{f} = \mathbf{f}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$. Denote $\boldsymbol{\zeta} = \boldsymbol{\Psi}^T\boldsymbol{\eta}$. Then, it is easy to check that

$$\boldsymbol{\Phi}^T(\mathbf{z} - \mathbf{f}) = \boldsymbol{\Psi}^T(\mathbf{y} - \mathbf{Q}\mathbf{f}) = \sigma\boldsymbol{\zeta}.$$

Following [14], by K-K-T condition, we derive that for any $\mathbf{t} \in \mathbb{R}^p$

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^T \boldsymbol{\Phi}^T (\mathbf{z} - \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}) &= \alpha \sum_{j=1}^p \nu_j |\widehat{\theta}_j| \\ \mathbf{t}^T \boldsymbol{\Phi}^T (\mathbf{z} - \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}) &\leq \alpha \sum_{j=1}^p \nu_j |t_j|, \end{aligned}$$

so that, subtracting the first line from the second, we obtain

$$(\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \boldsymbol{\Phi}\mathbf{t})^T (\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \mathbf{z}) \leq \alpha \sum_{j=1}^p \nu_j (|t_j| - |\widehat{\theta}_j|). \tag{3.25}$$

Then, (3.25) yields $(\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \boldsymbol{\Phi}\mathbf{t})^T (\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \boldsymbol{\Phi}\boldsymbol{\theta}) \leq \sigma(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T \boldsymbol{\zeta} + \alpha \sum_{j=1}^p \nu_j (|t_j| - |\widehat{\theta}_j|)$. Since for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ one has $\mathbf{v}^T \mathbf{u} = \frac{1}{2} \left[\|\mathbf{v}\|^2 + \|\mathbf{u}\|^2 - \|\mathbf{v} - \mathbf{u}\|^2 \right]$, choosing $\mathbf{v} = \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \boldsymbol{\Phi}\mathbf{t}$ and $\mathbf{u} = \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \boldsymbol{\Phi}\boldsymbol{\theta}$ for any $\mathbf{t} \in \mathbb{R}^p$ obtain

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|^2 + \|\boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|^2 \leq \|\mathbf{f}_{\mathbf{t}} - \mathbf{f}\|^2 + 2\sigma(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T \boldsymbol{\zeta} + 2\alpha \sum_{j=1}^p \nu_j (|t_j| - |\widehat{\theta}_j|). \tag{3.26}$$

By definition of $\boldsymbol{\zeta}$, for any $j = 1, \cdots, p$, one has $\zeta_j \sim \mathcal{N}(0, \nu_j^2)$. Hence, on the set

$$\Omega_0 = \left\{ \omega : \max_{1 \leq j \leq p} (\nu_j^{-1} |\zeta_j|) \leq \sqrt{2(\tau+1)\log p} \right\} \quad \text{with} \quad \mathbb{P}(\Omega_0) \geq 1 - 2p^{-\tau} \tag{3.27}$$

one obtains $|(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T \boldsymbol{\zeta}| \leq \sqrt{2(\tau+1)\log p} \sum_{j=1}^p \nu_j |\widehat{\theta}_j - t_j| = \alpha_0 \sum_{j=1}^p \nu_j |\widehat{\theta}_j - t_j|$. Combining the last inequality with (3.26) obtain that, for any $\alpha > 0$, on the set Ω_0,

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|^2 + \|\boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|^2 \leq \|\mathbf{f}_{\mathbf{t}} - \mathbf{f}\|^2 + 2\alpha \sum_{j=1}^p \nu_j (|t_j| - |\widehat{\theta}_j|) + 2\alpha_0 \sum_{j=1}^p \nu_j |\widehat{\theta}_j - t_j|. \tag{3.28}$$

Denote $\boldsymbol{\Omega}_1 = \{\omega : \lambda_{\min}(2s; \boldsymbol{\Phi}) \geq 1 - \delta\}$, $\boldsymbol{\Omega}_2 = \{\omega : |J_*| \leq s\}$ and $\boldsymbol{\Omega}_3 = \left\{\omega : \|\widehat{\boldsymbol{\theta}}\|_0 \leq s\right\}$. Choose $\mathbf{t}$ such that $\mathbf{f}_{\mathbf{t}} = \text{proj}_{\mathcal{L}_{J_*}} \mathbf{f} = \mathbf{f}_{\mathcal{L}_{J_*}}$ and note that $t_j = 0$ for $j \in J_*^c$. Then, due to $\alpha \geq \alpha_0$ and $||\widehat{\theta}_j - t_j| \leq |\widehat{\theta}_j| + |t_j|$, obtain

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|^2 + \|\boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|^2 \leq \|\mathbf{f}_{\mathbf{t}} - \mathbf{f}_{\mathcal{L}_{J_*}}\|^2 + 4\alpha \sum_{j \in J_*} \nu_j |\widehat{\theta}_j - t_j|. \tag{3.29}$$

Consider the set $\mathbf{\Omega} = \mathbf{\Omega}_0 \cap \mathbf{\Omega}_1 \cap \mathbf{\Omega}_2 \cap \mathbf{\Omega}_3$ and note that $\mathbb{P}(\mathbf{\Omega}) \geq 1 - 2p^{-\tau} - \epsilon_1 - \epsilon_2 - \epsilon_3$. If $\omega \in \mathbf{\Omega}$, then $\|\widehat{\boldsymbol{\theta}} - \mathbf{t}\|_0 \leq 2s$ and, hence,

$$4\alpha \sum_{j \in J_*} \nu_j \, |\widehat{\theta}_j - t_j| \leq 4\alpha \left(\sum_{j \in J_*} \nu_j^2 \right)^{1/2} \frac{\|\mathbf{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|}{\lambda_{\min}(2s; \mathbf{\Phi})} \leq \|\mathbf{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|^2 + \frac{4\alpha^2}{(1-\delta)^2} \sum_{j \in J_*} \nu_j^2.$$

Plugging the last inequality into (3.29) and recalling the definition of J_*, we derive (3.17).

Proof of Lemma 3.3.2. In cases a) and b), $\lambda_{\min}(m; \mathbf{\Phi}) \geq 1 - \delta$ is ensured by Theorem 5.65 of Vershynin [33]. In case c), note that entries of matrix $\mathbf{\Phi}$ are uncorrelated and, hence, are independent Gaussian variables due to

$$\mathrm{Cov}\,(\mathbf{\Phi}_{ik}, \mathbf{\Phi}_{jl}) = \frac{1}{c^2} \sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \mathbf{D}_{ir_1} \mathbf{D}_{jr_2} I(r_1 = r_2) I(k = l) = I(i = j) I(k = l).$$

Moreover, matrix $\mathbf{\Phi}$ has isotropic rows since

$$\mathrm{Cov}\,(\mathbf{\Phi}_{ih}, \mathbf{\Phi}_{jl}) = \frac{1}{c^2} \sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \mathbf{D}_{ir_1} \mathbf{D}_{jr_2} I(r_1 = r_2) I(h = l) = I(i = j) I(h = l).$$

Therefore, $\lambda_{\min}(m; \mathbf{\Phi}) \geq 1 - \delta$ by Theorem 5.65 of [33].

CHAPTER 4: SOLUTION OF ILL-POSED LINEAR INVERSE PROBLEMS WITH NON-GAUSSIAN NOISE USING OVERCOMPLETE DICTIONARIES

In this chapter, similar to Chapter 3 we address the linear ill-posed problem (1.1) where the observations are contaminated by non-Gaussian noise. We define this problem as,

$$y_i = r\lambda_i + \epsilon_i, \text{ where } \lambda_i = (\mathbf{Qf})_i, \text{ for } i = 1, 2, \cdots, n \tag{4.1}$$

where the observations y_i for $i = 1, 2, \cdots, n$ are independent and it has probability mass function that depend on the parameter $\lambda_i, i = 1, 2, \cdots, n$ and r is known. Also, here $\mathbf{f} \in \mathbb{R}^n$ is the function of interest, and, $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is a linear operator.

Note that, Var(y_i) is a function of λ_i. From (4.1) we have, $\epsilon_i = y_i - r\lambda_i$, for $i = 1, 2, \cdots, n$ which are deviations of the observations from the unobserved variables. In what follows, we assume that function $\lambda(x)$ is integrable, so that $\|\boldsymbol{\lambda}\|_1 = \frac{1}{n}\sum_{j=1}^{p} \lambda_i$ is finite.

Since, $\mathbb{E}(y_i) = r\lambda_i$ we have, $\mathbb{E}\epsilon_i = 0$. Alternatively, (4.1) can also be viewed as,

$$\mathbf{y} = \mathbf{Qf} + \boldsymbol{\epsilon} \text{ with } r\boldsymbol{\lambda} = \mathbf{Qf}, \tag{4.2}$$

where, and $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the vector with entries λ_i.

4.1 Construction of the Lasso Estimator

As discussed in Chapter 3, our idea is that we can find a good approximation of the function of interest $\mathbf{f}$, the true solution of the problem (4.2) by expanding $\mathbf{f}$ over the elements of an uninorm

basis $\{\boldsymbol{\varphi}_j(t)\}_{j=1}^p$, also known as dictionary elements in our context, as

$$f_t = \sum_{j=1}^{p} t_j \varphi_j,$$

where $\mathbf{t} \in \mathbb{R}^p$ with elements t_j, $j = 1, 2, \cdots, p$.

Proceeding similarly as in [28], let $\mathbf{f}_{\boldsymbol{\theta}}$ be the projection of the true solution $\mathbf{f}$ on the linear span of functions $\{\varphi_j, j \in \mathcal{P}\}$. We want to solve the following optimization problem:

$$\boldsymbol{\theta} = \arg\min_{\mathbf{t}} \|\mathbf{f} - \mathbf{f}_{\mathbf{t}}\|_2^2,$$

which is equivalent to,

$$\boldsymbol{\theta} = \arg\min_{\mathbf{t}} \left[\|\mathbf{f}_{\mathbf{t}}\|^2 - 2\sum_{j=1}^{p} \langle f, \varphi_j \rangle t_j \right]. \tag{4.3}$$

In order to estimate (4.3) we assume that the following condition holds for the dictionary elements $\boldsymbol{\varphi}_j$ for $j = 1, 2, \cdots, p$. :

Assumption A0: There exist $\boldsymbol{\psi}_j$ such that $\mathbf{Q}^* \boldsymbol{\psi}_j = \boldsymbol{\varphi}_j$ and $\nu_j = \|\boldsymbol{\psi}_j\|_\infty < \infty$ for $j = 1, 2, \cdots, p$.

Under the Assumption **A0** we can write,

$$\beta_j = \langle f, \varphi_j \rangle = \langle r\lambda, \psi_j \rangle = r \int_0^1 \psi_j(x)\lambda(x)dx, \tag{4.4}$$

Since our data is discrete, β_j in (4.4) can not be estimated by $\langle y, \psi_j \rangle$ directly. Instead we first use rectangular approximation of β_j:

$$\tilde{\beta}_j = \frac{r}{n} \sum_{i=1}^{n} \psi_j \left(\frac{i}{n} \right) \lambda_i, \quad j = 1, \cdots, p, \tag{4.5}$$

and then we estimate $\tilde{\beta}_j$ by

$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \psi_j \left(\frac{i}{n} \right) y_i, \quad j = 1, \cdots, p. \tag{4.6}$$

Finally, we find $\widehat{\boldsymbol{\theta}}$ as a solution of the following lasso optimization problem:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \left[\|\mathbf{f_t}\|^2 - 2\sum_{j=1}^{p} \widehat{\beta}_j t_j + \alpha \sum_{j=1}^{p} \nu_j |t_j| \right], \tag{4.7}$$

where α is a lasso penalty parameter. In order to reduce optimization problem (4.7) to familiar matrix formulation, we introduce matrix $\mathbf{\Phi}$ with elements $\Phi_{jk} = \langle \varphi_j, \varphi_k \rangle$ and vectors $\widehat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ with elements $\widehat{\beta}_j$, $\tilde{\beta}_j$ respectively for $j = 1, 2, \cdots, p$. Define matrices $\mathbf{W}$ by $\mathbf{W}^T\mathbf{W} = \mathbf{\Phi}$. Then (4.7) can be written as,

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \left[\|\mathbf{W}\mathbf{t} - \boldsymbol{\gamma}\|_2^2 + \alpha \|\Upsilon \mathbf{t}\|_1 \right] \quad \text{with} \quad \Upsilon = \text{diag}(\nu_1, \nu_2, \cdots, \nu_p). \tag{4.8}$$

Here, $\|\Upsilon\mathbf{t}\|_1$ is the weighted lasso penalty, α is the penalty parameter and

$$\boldsymbol{\gamma} = \left(\mathbf{W}^T\right)^{\dagger} \widehat{\boldsymbol{\beta}} = \left(\mathbf{W}\mathbf{W}^T\right)^{-1} \mathbf{W}\widehat{\boldsymbol{\beta}}. \tag{4.9}$$

Consider the vector $\boldsymbol{\eta} = \{\eta_1, \eta_2, \cdots, \eta_p\}^T$ in $\mathbb{R}^p$ with

$$\eta_j = \frac{1}{\nu_j} \left(\hat{\beta}_j - \tilde{\beta}_j \right), \tag{4.10}$$

such that,

$$\widehat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \Upsilon \boldsymbol{\eta},$$

The error of the lasso estimator (4.8) depends on the rate at which η_j decline as n grows. For this reason, in the next section, we examine behaviors of η_j for several discrete distributions.

4.2 Estimation of Functionals for Various Noise Distribution

Denote

$$\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n}\lambda_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{4.11}$$

and consider several noise scenarios.

4.2.1 Poisson Noise

Let us consider a particular case of (4.1) where the observations are independent Poisson variables defined as,

$$y_i \sim \text{ Poisson } (\lambda_i), \text{ for } i = 1, 2, \cdots, n. \tag{4.12}$$

In this case,

$$\mathbb{E}(y_i) = \lambda_i, \quad \text{Var}(\widehat{\beta}_j) = \frac{1}{n^2}\sum_{i=1}^{n}\psi_j^2\left(\frac{i}{n}\right)\lambda_i, \text{ for } i = 1, 2, \cdots, n,$$

and (4.5) and (4.6) hold with $r = 1$. Adopting Proposition 7 of [29] to our settings, we obtain for any $t > 0$

$$\mathbb{P}(|\hat{\beta}_j - \tilde{\beta}_j| \geq t) \;\leq\; 2\exp\left(-\frac{t^2}{\frac{2}{n^2}\sum_{i=1}^{n}\psi_j^2(\frac{i}{n})\lambda_i + \frac{2t\|\psi_j\|_\infty}{3n}}\right) \leq 2\exp\left(-\frac{t^2}{2\frac{\nu_j^2}{n}\bar{\lambda} + \frac{2t\nu_j}{3n}}\right) \tag{4.13}$$

The value of the threshold t however depends on the unknown quantity $\bar{\lambda}$. Nevertheless, the advantage of the formulation is that, when n is large enough, one can accurately estimate $\bar{\lambda}$ from the data. In particular, the following statement is valid:

Lemma 4.2.1. *Let y_i be independent Poisson variables with parameters λ_i, for $i = 1, 2, \cdots, n$ as defined in* (4.12)*. Then, for any $\tau > 0$*

$$\mathbb{P}\left(\bar{\lambda} \leq 12\tau \log p\, n^{-1} + 2\bar{y}\right) \geq 1 - p^{-\tau}, \tag{4.14}$$

where $\bar{\lambda}$ and $\bar{y}$ are defined in (4.11).

Combination of (4.13) and Lemma 4.2.1 yields the following result:

Lemma 4.2.2. *Let y_i be independent Poisson variables with parameters λ_i, for $i = 1, 2, \cdots, n$ as defined in* (4.12). *Then, for any $\tau > 0$*

$$\mathbb{P}\left(|\hat{\beta}_j - \tilde{\beta}_j| < \frac{2\nu_j}{\sqrt{n}}\left[\sqrt{2\bar{y}\tau\log p} + \sqrt{12}\cdot\frac{\tau\log p}{\sqrt{n}}\right]\right) \geq 1 - 3p^{-\tau}, \tag{4.15}$$

where $\tilde{\beta}_j$, $\widehat{\beta}_j$ and $\bar{y}$ are defined in (4.5), (4.6) *and* (4.11), *respectively.*

4.2.2 Binomial Noise

Let us consider another specific case of (4.1) where the observations are independent Binomial variables defined as,

$$y_i \sim \text{ Bin}(r, \lambda_i), \text{ for } i = 1, 2, \cdots, n. \tag{4.16}$$

In this case,

$$E(y_i) = r\lambda_i, \quad \text{Var}(y_i) = r\lambda_i(1-\lambda_i), \text{ for } i = 1, 2, \cdots, n.$$

Then, by deriving an upper bound for $\bar{\lambda}$ and plugging it into the Bernstein inequality, we obtain the following statement.

Lemma 4.2.3. *Let y_i be independent Binomial variables with parameters (r, λ_i), for $i = 1, 2, \cdots, n$ as defined in* (4.16). *Then, for any $\tau > 0$*

$$\mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| < \frac{2\nu_j}{\sqrt{n}}\left[\sqrt{2\bar{y}\tau\log p} + \sqrt{4 + \frac{8r}{3}}\cdot\frac{\tau\log p}{\sqrt{n}}\right]\right) \geq 1 - 3p^{-\tau}, \tag{4.17}$$

where $\tilde{\beta}_j$, $\widehat{\beta}_j$ and $\bar{y}$ are defined in (4.5), (4.6) *and* (4.11), *respectively.*

4.2.3 Chi-square Noise

Finally, we consider the particular case of (4.1) where the observations are independent Chi-squared variables defined as,

$$y_i \sim \ \chi^2(\lambda_i), \quad \text{for } i = 1, 2, \cdots, n. \tag{4.18}$$

In this case,

$$E(y_i) = \lambda_i, \quad \text{Var}(y_i) = 2\lambda_i.$$

Then, similar to the previous case, by deriving an upper bound for $\bar{\lambda}$ and plugging it into the Bernstein inequality, we obtain the following statement.

Lemma 4.2.4. *Let y_i be independent Chi-square variables with degree of freedom λ_i for $i = 1, 2, \cdots, n$ as defined in* (4.18)*. Then, for any $\tau > 0$,*

$$\mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| < \frac{2\nu_j}{\sqrt{n}}\left[\sqrt{2\bar{y}\tau\log p} + 8\cdot\frac{\tau\log p}{\sqrt{n}}\right]\right) \geq 1 - 3p^{-\tau}, \tag{4.19}$$

where $\tilde{\beta}_j$, $\widehat{\beta}_j$ and $\bar{y}$ are defined in (4.5)*,* (4.6) *and* (4.11)*, respectively.*

4.3 Oracle Inequalities for the Error

Theorem 4.3.1. *Let* f *be the true function and* f_θ *be its projection onto the linear span of the dictionary $\mathcal{L}_\mathcal{P}$. Consider solution of the weighted lasso problem* (4.8) *with $\mathbf{\Phi} = \mathbf{W}^T\mathbf{W}, \boldsymbol{\beta} = \mathbf{\Phi}\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\beta}}$ given by* (4.6) *and let,*

$$\aleph = \max_{1\leq j\leq p}\left[\frac{1}{\nu_j}\max_{x\in\mathcal{X}}|\frac{d}{dx}[r\lambda(x)\psi_j(x)]|\right], \tag{4.1}$$

and

$$\mathbb{P}\left(|\eta_j| < \frac{2}{\sqrt{n}}\left[\sqrt{2\bar{y}\tau\log p} + C_0\frac{\tau\log p}{\sqrt{n}}\right]\right) \geq 1 - 3p^{-\tau}, \tag{4.2}$$

where C_0 depends on the probability distribution of y_i and Let η_j and $\bar{y}$ be defined as in (4.10)*, and* (4.11) *respectively. Choose $\tau > 0$ and denote*

$$\alpha_0 = 2g_{\tau,p}(\bar{y}), \tag{4.3}$$

where,

$$g_{\tau,p}(x) = \frac{2}{\sqrt{n}}\sqrt{2x\tau\log p} + C_0\frac{\tau\log p}{\sqrt{n}}. \tag{4.4}$$

Then for any $\alpha \geq \alpha_0$ and,

$$n \geq \frac{\aleph}{2g_{\tau,p}(\bar{y})}, \tag{4.5}$$

with probability atleast $1 - 3p^{-\tau}$, one has

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 \leq \inf_{\mathbf{t}} \|\mathbf{f}_{\mathbf{t}} - \mathbf{f}\|_2^2 + 4\alpha\|\Upsilon\mathbf{t}\|_1. \tag{4.6}$$

Moreover, if assumption $(\mathbf{A}1)$ holds and $\alpha = \bar{\omega}\alpha_0$ where $\bar{\omega} \geq \frac{\mu+1}{\mu-1}$, then for any $\tau > 0$ with probability at least $1 - 2p^{-\tau}$, one has

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 \leq \inf_{t,J\subseteq\mathcal{P}} \left[\|\mathbf{f}_t - \mathbf{f}\|_2^2 + 4\alpha\|(\Upsilon\mathbf{t})_{J^c}\|_1 + \frac{(1+\bar{\omega}^2)}{\kappa^2(\mu,J)}\alpha_0^2\sum_{j\in J}\nu_j^2\right]. \tag{4.7}$$

Corollary 4.3.1.1.

- *If y_i are independent Poisson variables $i = 1, 2, \cdots, n$, Theorem 4.3.1 holds for $C_0 = \sqrt{12}$.*
- *If y_i are independent Binomial variables $i = 1, 2, \cdots, n$, Theorem 4.3.1 holds for $C_0 = \sqrt{4 + \frac{8r}{3}}$.*
- *If y_i are independent Chi-square variables $i = 1, 2, \cdots, n$, Theorem 4.3.1 holds for $C_0 = 8$.*

4.4 Simulations Studies

4.4.1 Simulation Setup

To evaluate the performance of the suggested methodology, we carried out a limited simulation study using various signals corrupted by different class of non-Gaussian noises. For our simulation we chose three sample sizes as $n = 128$, $n = 256$, and, $n = 512$ which are uniformly spaced over the interval $[0, 1]$. The simulation can also be extended to a general case of un-uniformly spaced interval $[0, T]$ with $T > 1$.

We first created a variety of well-known signals in the field of signal processing for our simulation like 'Wave', 'Parabolas,' and, 'Corner' as the test signals $\mathbf{f}$ on discrete time points as vectors in MATLAB. We then generated the $n \times n$ linear operator $\mathbf{Q}$ as Toeplitz matrix. The first row of the Toeplitz matrix $\mathbf{Q}$ is such that it has the first eight entries as ones and the remaining entries as zeros. Also, the first column of $\mathbf{Q}$ is such that it has first entry as one, next $\frac{n}{4} - 1$ entries as 0.5, and, the remaining $\frac{3n}{4}$ entries as zeros. The unobserved vector $\boldsymbol{\lambda}$ as in (4.2) is generated using $\mathbf{Q}$ and at last, we obtained the observed variables $\mathbf{y}$ as given in (4.1) and (4.2) by adding non-Gaussian noise. We verified the efficiency of our methodology for two different kinds of non-Gaussian noise as mentioned in section 4.2: 'Poisson' and 'Chi-square'.

We created our overcomplete dictionary which is rich enough containing it's columns as a variety of class of functions such as 'Step', 'Laguerre', 'Haar', 'Daubechies', 'Fourier' etc. so that it could represent test signals efficiently.

4.4.2 Implementation Details

We followed the procedure at the beginning of this chapter for the estimation of the function of interest $\mathbf{f}$. We first obtained matrix $\boldsymbol{\Psi}$ whose elements are the inverse images as the numerical so-

lution of the exact equation $\mathbf{Q}^T\mathbf{\Psi} = \mathbf{W}$ and then we obtained the vector $\boldsymbol{\gamma}$ with elements (4.9). For obtaining a solution of the optimization problem (4.8), we used function `mexLassoWeighted` defined in SPAMS MatLab toolbox (see [27]) Mairal (2014)). (4.8) is a weighted lasso problem with lasso parameter α and hence to optimize the accuracy of our estimator we first found the value of the lasso parameter α which optimizes the objective function in (4.8). In order to do that, we calculated $\alpha_{\max}$ as the value of α which is responsible for all the coefficients in the model to vanish. We then created a uniform grid of the values $\alpha_k = \alpha_{\max} * k/N$, $k = 1, \cdots, N$, of α with $N = 200$. Using α_k, $k = 1, \cdots, N$, we obtained a collection of estimators $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\alpha_k)$ and finally we obtained $\widehat{k}$, the most appropriate value of k so that $\hat{\alpha} = \alpha_{\widehat{k}}$ optimizes the objective function in (4.8). We estimated α as $\hat{\alpha} = \alpha_{\widehat{k}}$ in two different ways: one using the oracle value of α, and, using the estimated value of α. Oracle value of α is obtained as $\alpha_{oracle} = \alpha_{\max} * \widehat{k}_{oracle}/N$ using the value $\widehat{k}_{oracle}$ that guarantees the most accurate estimation of $\mathbf{f}$:

$$\widehat{k}_{oracle} = \arg\min_k \log \|\mathbf{f} - \mathbf{W}\widehat{\boldsymbol{\theta}}(\alpha_k)\|_2.$$

Unfortunately, since the true vector $\mathbf{f}$ is unavailable in real-life, we need to find the estimated value α and it is calculated similar to α_{oracle} as $\hat{\alpha}_{est} = \alpha_{\max} * \widehat{k}_{est}/N$ of α using

$$\widehat{k}_{oracle} = \arg\min_k \left[\log \|\mathbf{y} - \mathbf{W}\widehat{\boldsymbol{\theta}}(\alpha_k)\|_2 + p_0 * \log(n)/n\right].$$

where p_0 is the number of non zero coefficients in the model.

4.4.3 Results

We compared the proposed estimators of $\mathbf{f}$ based on the over-complete dictionary $\mathbf{W}$ with the $\mathbf{f}^{SVD}_{oracle}$, estimator based on the Singular Value Decomposition (SVD). In order to find $\hat{\mathbf{f}}^{SVD}_{oracle}$, we used the oracle number K_{oracle} of eigenbasis functions. We obtained K_{oracle} as the number of eigenbasis functions that minimizes the norm of difference between $\hat{\mathbf{f}}^{SVD}_{oracle}$ and the true function $\mathbf{f}$

which is unavailable in a real life setting. Figure 4.1, Figure 4.2 represents the simulation results for 'Wave' signal for Poisson and Chi-square noises. Similarly, Figure 4.3, Figure 4.4 represents the same for 'Parabolas' signal and Figure 4.5, Figure 4.6 represents the same for 'Corners' signal. Table 4.1 below compares the accuracies of the over-complete dictionary based estimators with $\hat{\mathbf{f}}^{SVD}_{oracle}$ for Poisson and Chi-square noise. We measured the precision of an estimator $\hat{\mathbf{f}}$ as $R(\hat{\mathbf{f}}) = \|\hat{\mathbf{f}} - \mathbf{f}\|_2 / \|\hat{\mathbf{f}}\|_2$, which is the relative error due to the estimation of $\mathbf{f}$ by $\hat{\mathbf{f}}$. Table 4.1 lists the relative errors averaged over 100 simulation runs (with the standard deviations listed in parentheses). We report the error of estimations with both the oracle and the estimated value of α, $\hat{\mathbf{f}}^{lasso}_{oracle}$, and $\hat{\mathbf{f}}^{lasso}_{cv}$ respectively.

From Table 4.1 it follows that the accuracy of over-complete dictionary based estimator is much higher (approximately 15% on average) than the SVD estimator for $n = 256$ and $n = 512$. For low sample size $n = 128$, $\hat{\mathbf{f}}^{SVD}_{oracle}$ has advantage over $\hat{\mathbf{f}}^{lasso}_{oracle}$, and $\hat{\mathbf{f}}^{lasso}_{cv}$. The advantage of $\hat{\mathbf{f}}^{lasso}_{oracle}$ over $\hat{\mathbf{f}}^{SVD}_{oracle}$ is more significant than that of $\hat{\mathbf{f}}^{lasso}_{cv}$ since the latter estimators loose accuracy because of suboptimal choices of the parameter α. Nevertheless, in the majority of cases, they still exhibit better precision than $\hat{\mathbf{f}}^{SVD}_{oracle}$ although this is not entirely fair comparison since $\hat{\mathbf{f}}^{SVD}_{oracle}$ is based on the oracle choice of parameter K. This is because the over-complete dictionaries provide a more sparse representation of $\mathbf{f}$.

Table 4.1: The average values of the errors $R(\hat{\mathbf{f}})$ evaluated over 100 simulation runs of the estimators for different signals under Poisson noise (standard deviations of the errors are listed in the parentheses).

Wave

	$\hat{\mathbf{f}}^{lasso}_{oracle}$	$\hat{\mathbf{f}}^{lasso}_{cv}$	$\hat{\mathbf{f}}^{SVD}_{oracle}$
$n = 128$ (Poisson)	0.2167 (0.0521)	0.2339 (0.0633)	0.1962 (0.0113)
$n = 128$ (Chi-square)	0.2663 (0.0636)	0.2848 (0.0661)	0.2091 (0.0147)
$n = 256$ (Poisson)	0.1781 (0.0339)	0.1937 (0.0346)	0.2133 (0.0121)
$n = 256$ (Chi-square)	0.2101 (0.0399)	0.2257 (0.0470)	0.2259 (0.0147)
$n = 512$ (Poisson)	0.1719 (0.0300)	0.1859 (0.0297)	0.2152 (0.0148)
$n = 512$ (Chi-square)	0.2036 (0.0301)	0.2155 (0.0314)	0.2326 (0.0186)

Parabolas

	$\hat{\mathbf{f}}^{lasso}_{oracle}$	$\hat{\mathbf{f}}^{lasso}_{cv}$	$\hat{\mathbf{f}}^{SVD}_{oracle}$
$n = 128$ (Poisson)	0.2166 (0.0479)	0.2378 (0.0458)	0.2013 (0.0210)
$n = 128$ (Chi-square)	0.2265 (0.0183)	0.2786 (0.0670)	0.2252 (0.0196)
$n = 256$ (Poisson)	0.1866 (0.0316)	0.1989 (0.0352)	0.2180 (0.0113)
$n = 256$ (Chi-square)	0.2207 (0.0225)	0.2365 (0.0393)	0.2359 (0.0192)
$n = 512$ (Poisson)	0.1850 (0.0245)	0.1993 (0.0324)	0.2125 (0.0185)
$n = 512$ (Chi-square)	0.2244 (0.0339)	0.2404 (0.0414)	0.2358 (0.0217)

Corners

	$\hat{\mathbf{f}}^{lasso}_{oracle}$	$\hat{\mathbf{f}}^{lasso}_{cv}$	$\hat{\mathbf{f}}^{SVD}_{oracle}$
$n = 128$ (Poisson)	0.1442 (0.0385)	0.1547 (0.0435)	0.1157 (0.0096)
$n = 128$ (Chi-square)	0.1883 (0.0562)	0.1976 (0.0590)	0.1242 (0.0116)
$n = 256$ (Poisson)	0.1208 (0.0196)	0.1294 (0.0199)	0.1480 (0.0042)
$n = 256$ (Chi-square)	0.1422 (0.0259)	0.1502 (0.0294)	0.1524 (0.0066)
$n = 512$ (Poisson)	0.1154 (0.0108)	0.1216 (0.0124)	0.1488 (0.0126)
$n = 512$ (Chi-square)	0.1377 (0.0181)	0.1444 (0.0194)	0.1677 (0.0164)

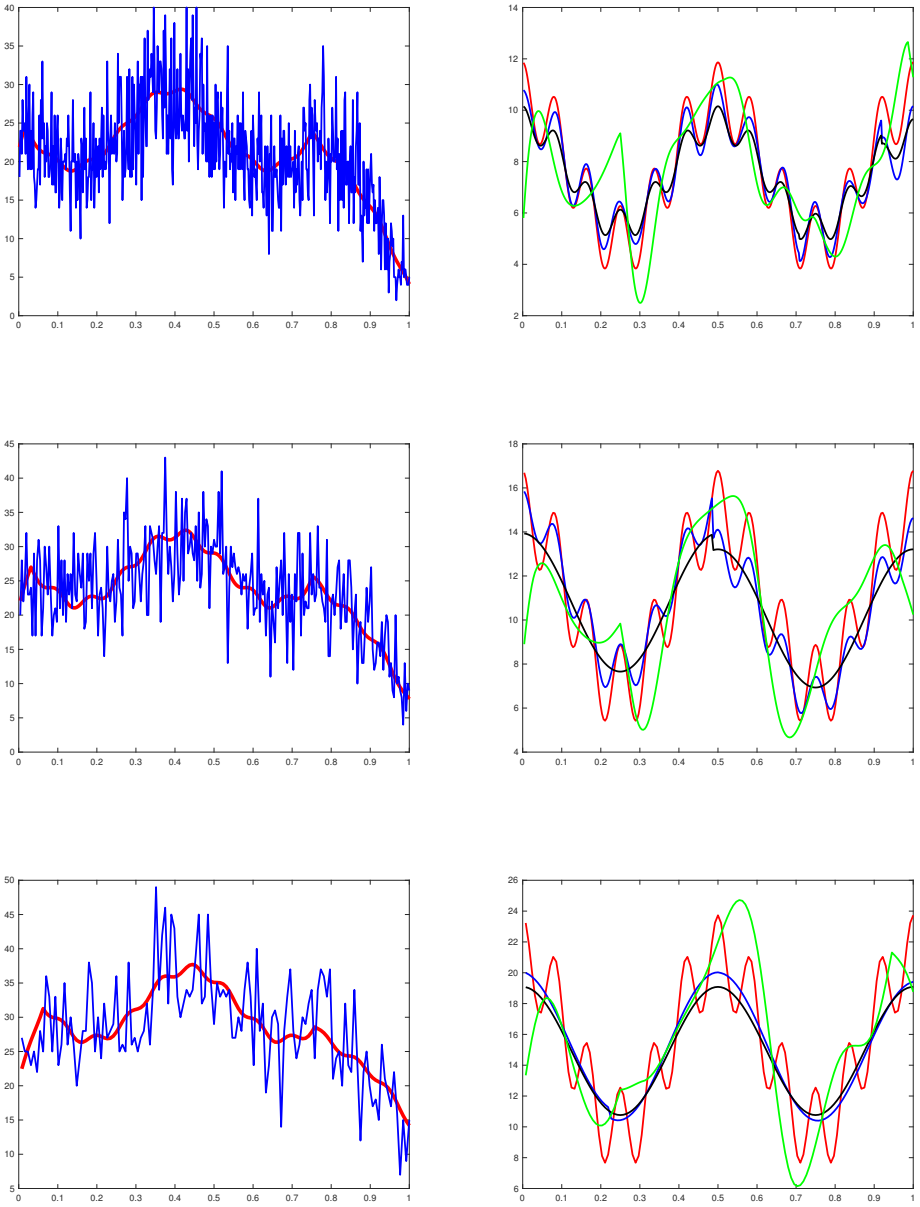

Figure 4.1: Simulation results for the 'Wave' signal under Poisson noise. Each figure on the left column contains the graphs of the unobserved signal $\mathbf{q}$ (red), and, the data $\mathbf{y}$ (blue). Each figures on the right column are the graphs of the true signal $\mathbf{f}$ (red), estimated signals $\mathbf{f}_{oracle}^{lasso}$ (blue), $\mathbf{f}_{cv}^{lasso}$ (black), and, $\mathbf{f}_{oracle}^{svd}$ (green). The figures in the first, second and third row corresponds to $n = 512$, $n = 256$, and, $n = 128$ respectively.

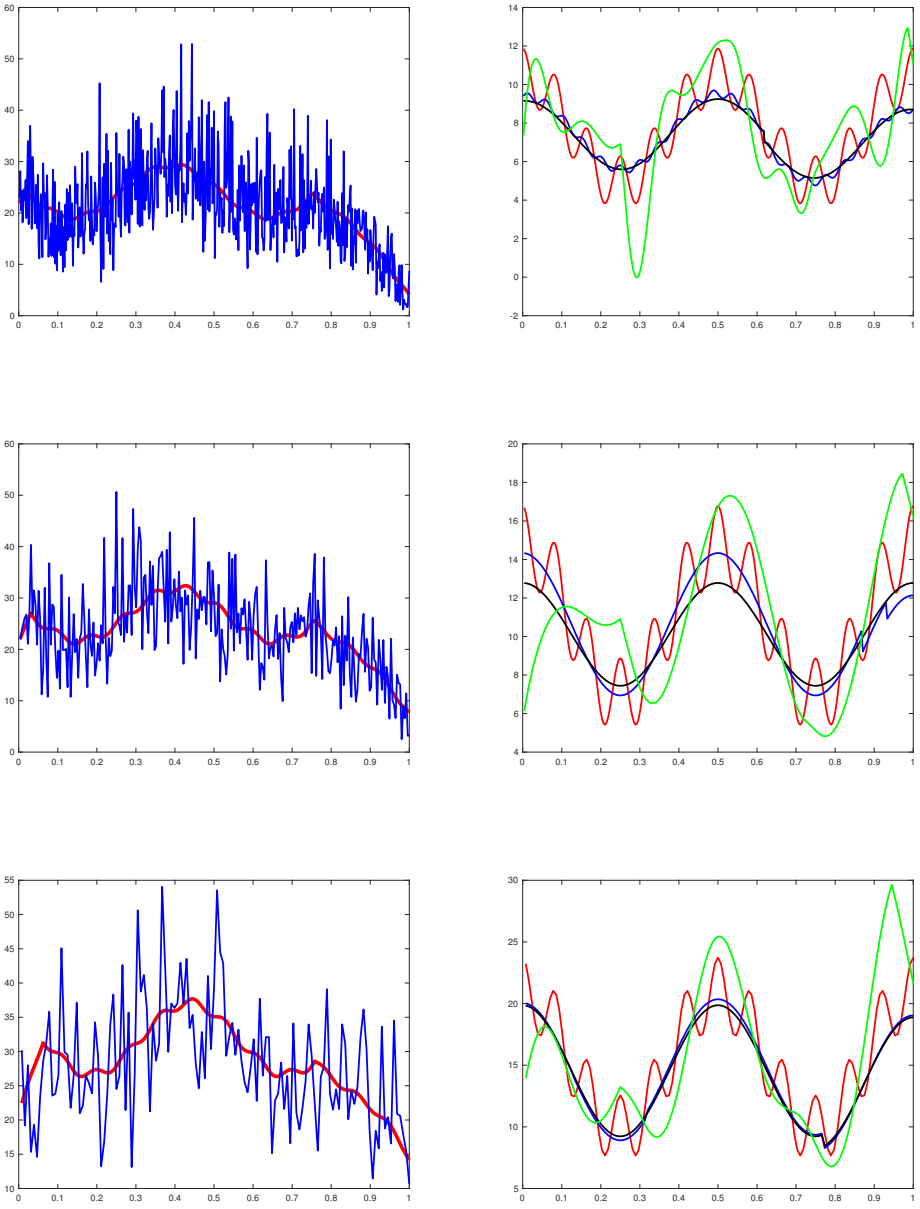

Figure 4.2: Simulation results for the 'Wave' signal under Chi-square noise. Each figure on the left column contains the graphs of the unobserved signal $\mathbf{q}$ (red), and, the data $\mathbf{y}$ (blue). Each figures on the right column are the graphs of the true signal $\mathbf{f}$ (red), estimated signals $\mathbf{f}_{oracle}^{lasso}$ (blue), $\mathbf{f}_{cv}^{lasso}$ (black), and, $\mathbf{f}_{oracle}^{svd}$ (green). The figures in the first, second and third row corresponds to $n = 512$, $n = 256$, and, $n = 128$ respectively.

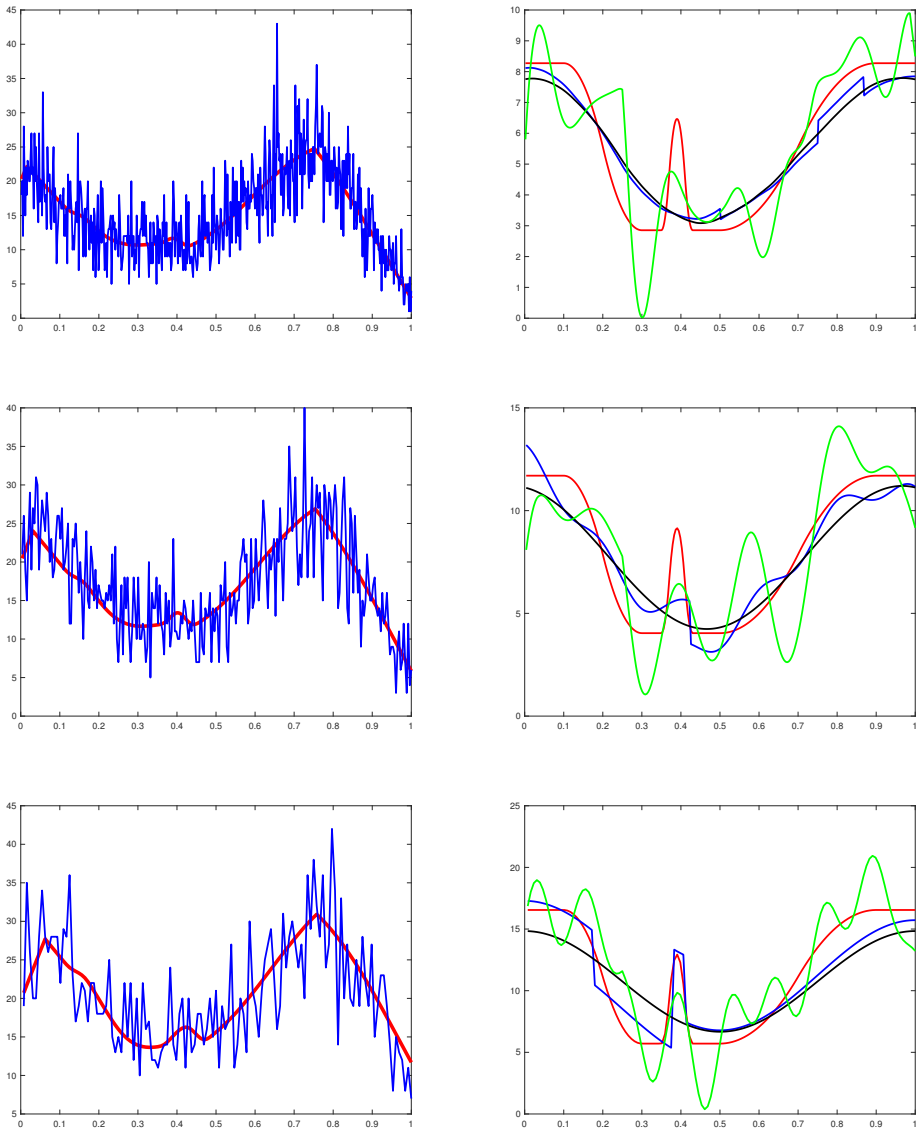

Figure 4.3: Simulation results for the 'Parabolas' signal under Poisson noise. Each figure on the left column contains the graphs of the unobserved signal $\mathbf{q}$ (red), and, the data $\mathbf{y}$ (blue). Each figures on the right column are the graphs of the true signal $\mathbf{f}$ (red), estimated signals $\mathbf{f}_{oracle}^{lasso}$ (blue), $\mathbf{f}_{cv}^{lasso}$ (black), and, $\mathbf{f}_{oracle}^{svd}$ (green). The figures in the first, second and third row corresponds to $n = 512$, $n = 256$, and, $n = 128$ respectively.

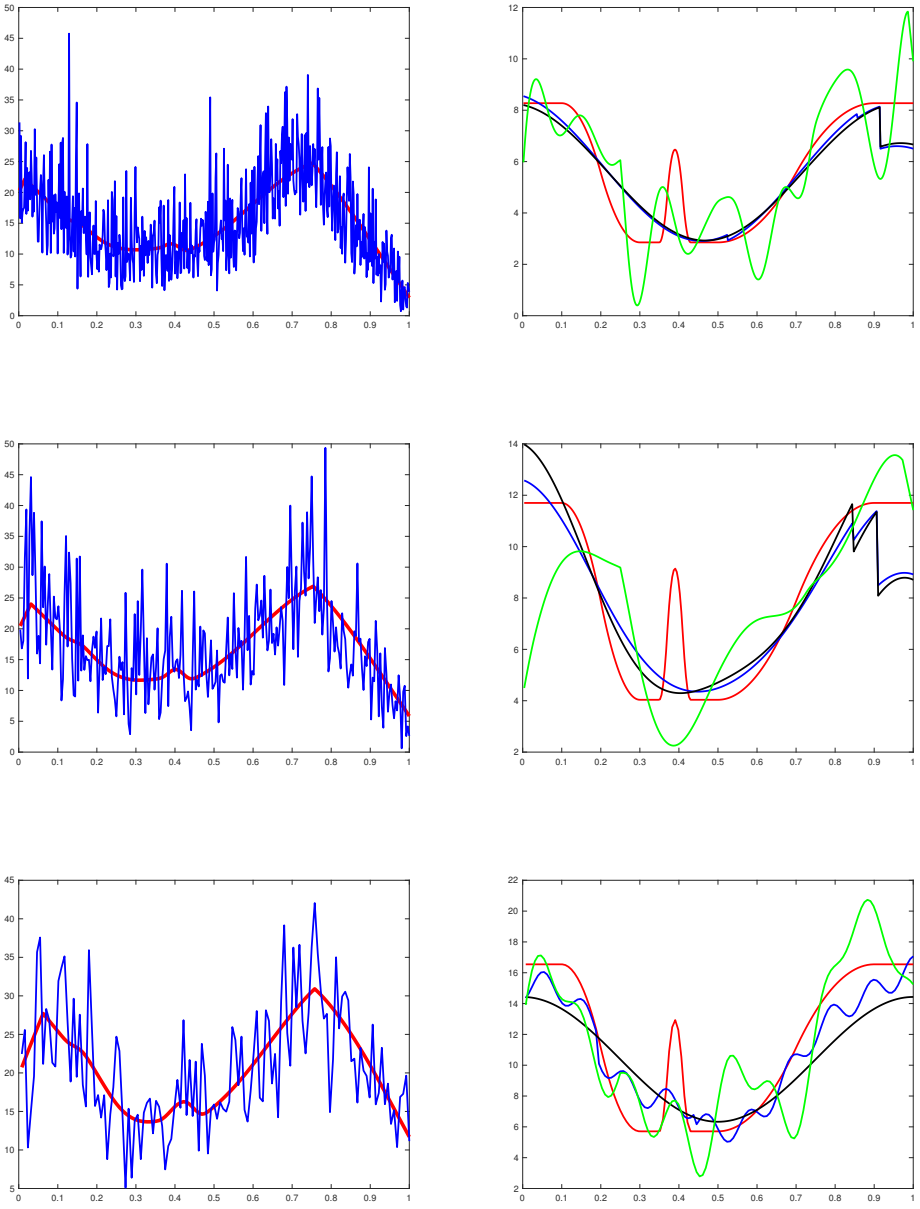

Figure 4.4: Simulation results for the 'Parabolas' signal under Chi-square noise. Each figure on the left column contains the graphs of the unobserved signal $\mathbf{q}$ (red), and, the data $\mathbf{y}$ (blue). Each figures on the right column are the graphs of the true signal $\mathbf{f}$ (red), estimated signals $\mathbf{f}_{oracle}^{lasso}$ (blue), $\mathbf{f}_{cv}^{lasso}$ (black), and, $\mathbf{f}_{oracle}^{svd}$ (green). The figures in the first, second and third row corresponds to $n = 512$, $n = 256$, and, $n = 128$ respectively.

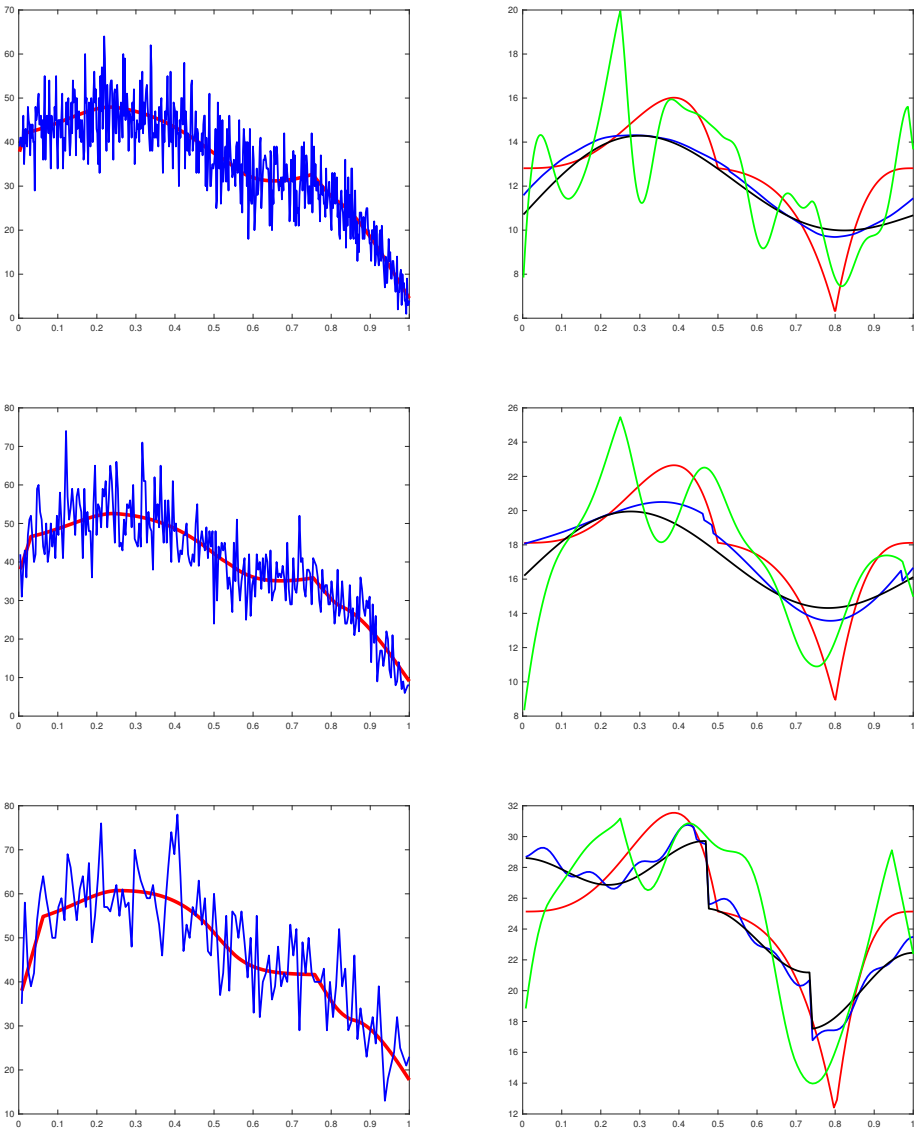

Figure 4.5: Simulation results for the 'Corners' signal under Poisson noise. Each figure on the left column contains the graphs of the unobserved signal $\mathbf{q}$ (red), and, the data $\mathbf{y}$ (blue). Each figures on the right column are the graphs of the true signal $\mathbf{f}$ (red), estimated signals $\mathbf{f}^{lasso}_{oracle}$ (blue), $\mathbf{f}^{lasso}_{cv}$ (black), and, $\mathbf{f}^{svd}_{oracle}$ (green). The figures in the first, second and third row corresponds to $n = 512$, $n = 256$, and, $n = 128$ respectively.

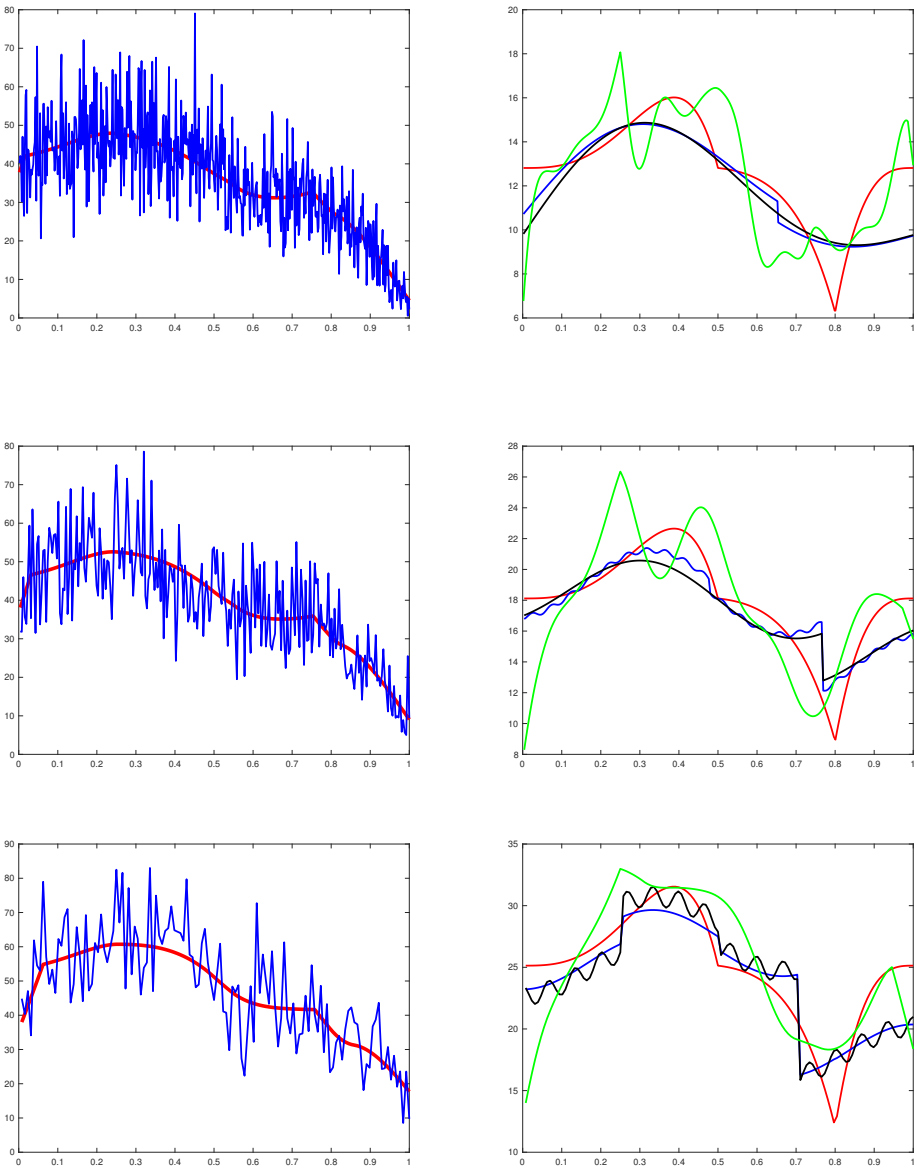

Figure 4.6: Simulation results for the 'Corners' signal under Chi-square noise. Each figure on the left column contains the graphs of the unobserved signal $\mathbf{q}$ (red), and, the data $\mathbf{y}$ (blue). Each figures on the right column are the graphs of the true signal $\mathbf{f}$ (red), estimated signals $\mathbf{f}_{oracle}^{lasso}$ (blue), $\mathbf{f}_{cv}^{lasso}$ (black), and, $\mathbf{f}_{oracle}^{svd}$ (green). The figures in the first, second and third row corresponds to $n = 512$, $n = 256$, and, $n = 128$ respectively.

4.5 Proofs

4.5.1 Proofs of the Lemmas

Proof of Lemma 4.2.1. Using Theorem 2.1 in [25] we obtain,

$$\mathbb{P}\left(\sum_{i=1}^{n}(y_i-\lambda_i)<-x\right)\leq\exp\left(-\frac{x^2}{2\sum_{i=1}^{n}\text{Var}(y_i)+2x}\right)=\exp\left(-\frac{x^2}{2n\bar{\lambda}+2x}\right)$$

Therefore,

$$\mathbb{P}\left(\bar{y}-\bar{\lambda}<-x\right)\leq\exp\left(-\frac{n^2x^2}{2n\bar{\lambda}+2nx}\right)=\exp\left(-\frac{x^2}{2\frac{\bar{\lambda}}{n}+2\frac{x}{n}}\right).$$

Equivalently,

$$\mathbb{P}\left(\bar{\lambda}<\bar{y}+x\right)\geq1-\exp\left(-\frac{n^2x^2}{2n\bar{\lambda}+2nx}\right)=1-\exp\left(-\frac{x^2}{2\frac{\bar{\lambda}}{n}+2\frac{x}{n}}\right).$$

Now,

$$\mathbb{P}\left(\bar{\lambda}<\bar{y}+x\right)\geq1-p^{-\tau},\tag{4.1}$$

provided,

$$\frac{x^2}{2\frac{\bar{\lambda}}{n}+2\frac{x}{n}}\geq\tau\log p,$$

which holds if,

$$\frac{x^2}{4\frac{\bar{\lambda}}{n}}\geq\tau\log p,\text{ and, }\frac{x^2}{4\frac{x}{n}}\geq\tau\log p,$$

Solving both the above inequalities for x we obtain,

$$x\geq\max\left(\sqrt{\frac{4\tau\log p\bar{\lambda}}{n}},\frac{4\tau\log p}{n}\right)$$

Therefore,

$$x\geq\sqrt{\frac{4\tau\log p\bar{\lambda}}{n}}+\frac{4\tau\log p}{n}$$

guarantees (4.1). Hence,

$$\mathbb{P}\left(\bar{\lambda} < \bar{y} + \sqrt{\frac{4\tau \log p \bar{\lambda}}{n}} + \frac{4\tau \log p}{n}\right) \geq 1 - p^{-\tau}, \tag{4.2}$$

Now solving the inequality,

$$\bar{\lambda} < \bar{y} + \sqrt{\frac{4\tau \log p \bar{\lambda}}{n}} + \frac{4\tau \log p}{n}$$

we get,

$$\bar{\lambda} < 2\bar{y} + \frac{12\tau \log p}{n}.$$

Hence, (4.14) follows.

Proof of Lemma 4.2.2. Note that it follows from (4.13) that

$$\mathbb{P}(|\hat{\beta}_j - \tilde{\beta}_j| \geq t) \leq 2p^{-\tau}$$

provided

$$\frac{t^2}{2\frac{\nu_j^2}{n}\bar{\lambda} + \frac{2t\nu_j}{3n}} \geq \tau \log p,$$

which is guaranteed by,

$$\frac{t^2}{4\frac{\nu_j^2}{n}\bar{\lambda}} \geq \tau \log p, \text{ and, } \frac{t^2}{\frac{4t\nu_j}{3n}} \geq \tau \log p.$$

Solving both inequalities for t, we obtain,

$$t \geq \frac{2\nu_j}{\sqrt{n}} \max\left\{\frac{2\tau \log p}{3\sqrt{n}}, \sqrt{\bar{\lambda}\tau \log p}\right\}. \tag{4.3}$$

Applying Lemma 4.2.1 and using (4.3), we rewrite (4.13) as

$$\mathbb{P}\left(|\hat{\beta}_j - \tilde{\beta}_j| \geq \frac{2\nu_j}{\sqrt{n}} \max\left\{\frac{2\tau \log p}{3\sqrt{n}}, \sqrt{\left[2\bar{y} + \frac{12\tau \log p}{n}\right]\tau \log p}\right\}\right) < 3p^{-\tau}.$$

Since, the second term in the maximum dominates the first term, we obtain (4.15).

Proof of Lemma 4.2.3. The proof of the statement relies on the following lemma proved in Section 4.5.3.

Lemma 4.5.1. *Let y_i be independent Binomial variables with parameters (r, λ_i) defined as in* (4.16)*. Then, for any $\tau > 0$*

$$\mathbb{P}\left(r\bar{\lambda} < 2\bar{y} + \frac{4+8r}{3} \cdot \frac{\tau \log p}{n}\right) \geq 1 - p^{-\tau}, \tag{4.4}$$

where $\bar{\lambda}$ and $\bar{y}$ defined in (4.11).

Note that

$$\widehat{\beta}_j - \tilde{\beta}_j = \sum_{i=1}^{n} \frac{1}{n} \psi_j \left(\frac{i}{n}\right) (y_i - r\lambda_i).$$

Let,

$$z_i = \frac{1}{n} \psi_j \left(\frac{i}{n}\right) (y_i - r\lambda_i).$$

Therefore,

$$\begin{aligned}
\mathbb{E}(z_i) = 0, \text{ and, } \sum_{i=1}^{n} \text{Var}(z_i) = \frac{1}{n^2} \sum_{i=1}^{n} \psi_j^2 \left(\frac{i}{n}\right) \text{Var}(y_i) &= \frac{1}{n^2} \sum_{i=1}^{n} \psi_j^2 \left(\frac{i}{n}\right) r\lambda_i(1-\lambda_i) \\
&\leq \frac{r\|\psi_j\|_\infty^2}{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \lambda_i(1-\lambda_i) \\
&\leq \frac{r\|\psi_j\|_\infty^2}{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \lambda_i \\
&\leq \frac{r\nu_j^2}{n} \bar{\lambda}.
\end{aligned}$$

Also, since $0 \leq y_i \leq r$ for $i = 1, 2, \cdots, n$,

$$\|z_i\|_\infty = \frac{1}{n}|\psi_j\left(\frac{i}{n}\right) \cdot (y_i - r\lambda_i)| \leq \frac{1}{n}\|\psi_j\|_\infty \cdot r \leq \frac{r\nu_j}{n}.$$

Therefore, using Bernstein inequality we derive,

$$\mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| > t\right) \leq 2\exp\left(-\frac{\frac{t^2}{2}}{\frac{r\nu_j^2}{n}\bar{\lambda} + \frac{r\nu_j t}{3n}}\right) \leq 2p^{-\tau}, \tag{4.5}$$

provided,

$$\frac{t^2}{2\frac{r\nu_j^2}{n}\bar{\lambda} + \frac{2rt\nu_j}{3n}} \geq \tau \log p,$$

which holds if,

$$\frac{t^2}{4r\frac{\nu_j^2}{n}\bar{\lambda}} \geq \tau \log p, \text{ and, } \frac{t^2}{\frac{4rt\nu_j}{3n}} \geq \tau \log p.$$

Solving both inequalities for t, we obtain,

$$t \geq \frac{2\nu_j}{\sqrt{n}} \max\left\{\frac{2r\tau \log p}{3\sqrt{n}}, \sqrt{r\bar{\lambda}\tau \log p}\right\}. \tag{4.6}$$

Using Lemma 4.5.1 and (4.6), we rewrite (4.5)

$$\mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| < \frac{2\nu_j}{\sqrt{n}} \max\left\{\frac{2r\tau \log p}{3\sqrt{n}}, \sqrt{\tau \log p\left[2\bar{y} + \frac{4+8r}{3} \cdot \frac{\tau \log p}{n}\right]}\right\}\right) \geq 1 - 3p^{-\tau}.$$

Since, the second term in maximum dominates the first term, we obtain

$$\mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| < \frac{2\nu_j}{\sqrt{n}}\sqrt{\tau \log p\left[2\bar{y} + \frac{4+8r}{3} \cdot \frac{\tau \log p}{n}\right]}\right) \geq 1 - 3p^{-\tau}.$$

and therefore, (4.17) holds.

Proof of Lemma 4.2.4. The proof of the statement relies on the following lemmas proved in Section 4.5.3.

Lemma 4.5.2. *Let $X \sim \chi^2(k)$. Then, X is sub-exponential with parameters $(4k, 4)$.*

Lemma 4.5.3. *Let X_is are independent with $X_i \sim$ Sub-exp$(4c_i, 4)$, where c_is are constants for $i = 1, 2, \cdots, n$. Then,*

$$\sum_{i=1}^{n} a_i y_i \sim \text{ Sub-exp}(4 \max_i |a_i|^2 \sum_{i=1}^{n} c_i, 4 \max_i |a_i|).$$

Lemma 4.5.4. *Let y_i be independent Chi-square variables with degree of freedom λ_i for $i = 1, 2, \cdots, n$ defined in* (4.18)*. Then, for any $\tau > 0$,*

$$\mathbb{P}\left(\bar{\lambda} < \bar{y} + \frac{8\tau \log p}{n} + \frac{\sqrt{8\tau \log p}}{n}\right) \geq 1 - p^{-\tau}, \tag{4.7}$$

where $\bar{\lambda}$ and $\bar{y}$ defined in (4.11)*.*

Using lemma 4.5.3 we can conclude,

$$\sum_{i=1}^{n} \psi_j(\frac{i}{n}) y_i \sim \text{ Sub-exp}(4n\|\psi_j\|_\infty^2 \bar{\lambda}, 4\|\psi_j\|_\infty).$$

Now, using Bartlett's lecture we get,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \psi_j(\frac{i}{n})(y_i - \mathbb{E}(y_i))\right| \geq t\right) \leq 2 \begin{cases} \exp\left(-\frac{t^2}{8n\|\psi_j\|_\infty^2 \bar{\lambda}}\right), & \text{if } 0 \leq t \leq n\|\psi_j\|_\infty \bar{\lambda}, \\ \exp\left(-\frac{t}{8\|\psi_j\|_\infty}\right), & \text{if } t > n\|\psi_j\|_\infty \bar{\lambda}, \end{cases}$$

or equivalently,

$$\begin{aligned} \mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| \geq t\right) &\leq 2 \begin{cases} \exp\left(-\frac{t^2}{8\bar{\lambda}\frac{\nu_j^2}{n}}\right), & \text{if } 0 \leq t \leq \nu_j \bar{\lambda} \\ \exp\left(-\frac{t}{8\frac{\nu_j}{n}}\right), & \text{if } t > \nu_j \bar{\lambda}. \end{cases} \\ &\leq 2p^{-\tau}, \end{aligned}$$

provided,

$$\frac{t^2}{\frac{8\bar{\lambda}\nu_j^2}{n}} \geq \tau \log p, \text{ and, } \frac{t}{8\frac{\nu_j}{n}} \geq \tau \log p.$$

Solving both inequalities for t we obtain,

$$t \geq \frac{2\nu_j}{\sqrt{n}} \max\left\{\sqrt{2\bar{\lambda}\tau \log p}, \frac{4\tau \log p}{\sqrt{n}}\right\},$$

Therefore, we can consider,

$$t \geq \frac{2\nu_j}{\sqrt{n}}\left[\sqrt{2\bar{\lambda}\tau \log p} + \frac{4\tau \log p}{\sqrt{n}}\right].$$

Since, we don't know $\bar{\lambda}$, in order to bound $\bar{\lambda}$ with high probability we use Lemma 4.5.4 and obtain

$$\mathbb{P}\left(|\widehat{\beta}_j - \tilde{\beta}_j| < \frac{2\nu_j}{\sqrt{n}}\left[\sqrt{\left(2\bar{y} + \frac{16\tau \log p}{n} + \frac{2\sqrt{8\tau \log p}}{n}\right)\tau \log p} + \frac{4\tau \log p}{\sqrt{n}}\right]\right) \geq 1 - 3p^{-\tau},$$

and therefore (4.19) holds.

4.5.2 Proofs of Theorems

Proof of Theorem 4.3.1 By our construction,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \Upsilon\boldsymbol{\eta} + \mathbf{h},$$

where $\boldsymbol{\eta}$, $\mathbf{h} \in \mathbb{R}^p$, $\mathbf{h}$ is a non-random vector with entries h_j such that $\mathbf{h} = \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Now,

$$h_j = \frac{r}{n}\sum_{i=1}^{n} \lambda(\frac{i}{n})\psi_j(\frac{i}{n}) - \int r\lambda(x)\psi_j(x)dx,$$

where, h_j is the error due to the rectangular approximation, so that

$$|h_j| \leq \frac{\nu_j \aleph}{2n} \tag{4.8}$$

where $\aleph$ is defined in (4.1). Following Dalalyan, Hebiri and Lederer (2014), by KKT condition, we derive for any $t \in \mathbb{R}^p$,

$$\widehat{\boldsymbol{\theta}}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}) = \alpha\|\Upsilon\widehat{\boldsymbol{\theta}}\|_1,$$

$$\mathbf{t}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}) \leq \alpha\|\Upsilon\mathbf{t}\|_1,$$

so that, subtracting the first line from the second, we obtain

$$(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T(\Phi\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\beta}}) \leq \alpha\left(\|\Upsilon t\|_1 - \|\Upsilon\widehat{\theta}\|_1\right).$$

Since $\boldsymbol{\Phi\theta} = \boldsymbol{\beta}$, the above equation yields

$$(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\Phi(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\Upsilon\boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\mathbf{h} + \alpha\left(\|\Upsilon t\|_1 - \|\Upsilon\widehat{\theta}\|_1\right).$$

Since for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ one has

$$\mathbf{v}^T\boldsymbol{\Phi}\mathbf{u} = \frac{1}{2}[\mathbf{v}^T\boldsymbol{\Phi}\mathbf{v} + \mathbf{u}^T\boldsymbol{\Phi}\mathbf{u} - (\mathbf{v} - \mathbf{u})^T\boldsymbol{\Phi}(\mathbf{v} - \mathbf{u})],$$

choosing $\mathbf{v} = \widehat{\boldsymbol{\theta}} - \mathbf{t}$ and $\mathbf{u} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ and observing that for any $\mathbf{t}$ (and, in particular, for $\mathbf{t} = \widehat{\boldsymbol{\theta}}$),

$$\|\mathbf{f}_\mathbf{t} - \mathbf{f}\|_2^2 = (\mathbf{t} - \boldsymbol{\theta})^T\Phi(\mathbf{t} - \boldsymbol{\theta}) + \|\mathbf{f}_{\boldsymbol{\theta}} - \mathbf{f}\|_2^2,$$

for any $\mathbf{t} \in \mathbb{R}^p$, one obtains

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\Phi(\widehat{\boldsymbol{\theta}} - \mathbf{t}) \leq \|\mathbf{f}_t - \mathbf{f}\|_2^2 + 2\left[(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\Upsilon\boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\mathbf{h}\right] + 2\alpha\left(\|\Upsilon\mathbf{t}\|_1 - \|\Upsilon\widehat{\boldsymbol{\theta}}\|_1\right). \tag{4.9}$$

Also, observe that on the set

$$\Omega = \left\{\omega : \max_{1\leq j\leq p}|\eta_j| < \frac{2}{\sqrt{n}}\left[\sqrt{2\bar{y}\tau\log p} + C_0\frac{\tau\log p}{\sqrt{n}}\right]\right\},$$

with $P(\Omega) \geq 1 - 3p^{-\tau}$ one has,

$$\begin{aligned} & |(\widehat{\boldsymbol{\theta}} - t)^T \Upsilon \boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T \mathbf{h}| \\ < & \left[\frac{2}{\sqrt{n}} \left(\sqrt{2\bar{y}\tau \log p} + C_0 \frac{\tau \log p}{\sqrt{n}} \right) + \max_{1 \leq j \leq p} \frac{|h_j|}{\nu_j} \right] \|\Upsilon(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1 \\ \leq & \left[g_{\tau,p}(\bar{y}) + \frac{\aleph}{2n} \right] \|\Upsilon(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1, \end{aligned}$$

where the last inequality is obtained using (4.8) and (4.4). Now, let n satisfies (4.5). Then,

$$\frac{\aleph}{2n} \leq g_{\tau,p}(\bar{y}).$$

and, with probability $1 - 3p^{-\tau}$, one has,

$$\begin{aligned} |(\widehat{\boldsymbol{\theta}} - t)^T \Upsilon \boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T \mathbf{h}| & < 2g_{\tau,p}(\bar{y}) \|\Upsilon(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1 \\ & = \alpha_0 \|\Upsilon(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1, \end{aligned}$$

where α_0 is defined in (4.3).

Combining the last inequality with (4.9), obtain that for any $\alpha > 0$, on the set Ω,

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T \boldsymbol{\Phi} (\widehat{\boldsymbol{\theta}} - \mathbf{t}) \leq \|\mathbf{f}_{\mathbf{t}} - \mathbf{f}\|_2^2 + 2\alpha \left(\|\Upsilon \mathbf{t}\|_1 - \|\Upsilon \widehat{\boldsymbol{\theta}}\|_1 \right) + 2\alpha_0 \left(\|\Upsilon(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1 \right). \quad (4.10)$$

Application of inequality

$$\|\Upsilon(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1 \leq (\|\Upsilon \mathbf{t}\|_1 - \|\Upsilon \widehat{\boldsymbol{\theta}}\|_1)$$

combined with $\alpha \geq \alpha_0$ completes the proof of the inequality (4.6). In order to proof the inequality (4.7), denote $d = \widehat{\boldsymbol{\theta}} - \mathbf{t}$ and observe that, due to

$$|t_j| - |\widehat{\theta}_j| \leq |\widehat{\theta}_j - t_j| \text{ and } |\widehat{\theta}_j| \geq |\widehat{\theta}_j - t_j| - |t_j|,$$

inequality (4.10) implies that, for any set $\mathcal{J} \subseteq \mathcal{P}$, one obtains

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}} - \mathbf{f}\|_2^2 + \mathbf{d}^T \boldsymbol{\Phi} \mathbf{d} \leq \|\mathbf{f}_{\mathbf{t}} - \mathbf{f}\|_2^2 + 4\alpha \|(\Upsilon \mathbf{t})_{J^c}\|_1 + 2(\alpha + \alpha_0) \|(\Upsilon \mathbf{d})_J\|_1 - 2(\alpha - \alpha_0) \|(\Upsilon \mathbf{d})_{J^c}\|_1. \quad (4.11)$$

Now, we consider two possibilities. If

$$(\alpha+\alpha_0)\|(\Upsilon\mathbf{d})_J\|_1 \leq (\alpha-\alpha_0)\|(\Upsilon\mathbf{d})_{J^c}\|_1,$$

then

$$\|\mathbf{f}_{\widehat{\boldsymbol{\theta}}}-\mathbf{f}\|_2^2+\mathbf{d}^T\boldsymbol{\Phi}\mathbf{d} \leq \|\mathbf{f}_{\mathbf{t}}-\mathbf{f}\|_2^2+4\alpha\|(\Upsilon\mathbf{t})_{J^c}\|_1$$

and (4.7) is valid. Otherwise, since $\alpha=\bar{\omega}\alpha_0$ with $\bar{\omega}\geq\frac{\mu+1}{\mu-1}$ implies that $\mu\geq\frac{\alpha+\alpha_0}{\alpha-\alpha_0}$, one has $d\in\mathcal{J}(\mu,J)$. Therefore, due to compatibility condition (**??**) and inequality $2ab\leq a^2+b^2$, one derives

$$\begin{aligned}(\alpha+\alpha_0)\|(\Upsilon\mathbf{d})_J\|_1 &\leq 2(\alpha+\alpha_0)\sqrt{\mathrm{Tr}(\Upsilon_J^2)\mathbf{d}^T\boldsymbol{\Phi}\mathbf{d}}/\kappa(\mu,J)\\ &\leq \mathbf{d}^T\boldsymbol{\Phi}\mathbf{d}+(\alpha+\alpha_0)^2\mathrm{Tr}(\Upsilon_J^2)/\kappa^2(\mu,J).\end{aligned}$$

Plugging the later into (4.11) and using $\alpha=\bar{\omega}\alpha_0$, obtain that (4.7) holds for any t.

4.5.3 Proofs of Auxiliary Statements

Proof of Lemma 4.5.1. Let $w_i=y_i-r\lambda_i$ for $i=1,2,\cdots,n$. Then,

$$\mathbb{E}(w_i)=0, \text{ and, } \mathrm{Var}(w_i)=\mathrm{Var}(y_i)\leq r\lambda_i, \text{ for } i=1,2,\cdots,n.$$

Also, $|w_i|\leq|y_i-r\lambda_i|\leq y_i\leq r$. Therefore, we get the following probability bound using Bernstein inequality,

$$\mathbb{P}(\sum_{i=1}^n r\lambda_i-\sum_{i=1}^n y_i\geq t)<\exp\left(-\frac{\frac{t^2}{2}}{\sum_{i=1}^n r\lambda_i+\frac{1}{3}rt}\right),$$

Therefore,

$$\mathbb{P}(r\bar{\lambda}-\bar{y}\geq t)<\exp\left(-\frac{\frac{nt^2}{2}}{r\bar{\lambda}+\frac{1}{3}rt}\right),$$

which is equivalent to,

$$\mathbb{P}(r\bar{\lambda} < \bar{y} + t) \geq 1 - p^{-\tau},$$

provided,

$$\frac{\frac{nt^2}{2}}{r\bar{\lambda} + \frac{1}{3}rt} \geq \tau \log p.$$

which holds if,

$$\frac{nt^2}{4r\bar{\lambda}} \geq \tau \log p \text{ and, } \frac{nt^2}{\frac{4}{3}rt} \geq \tau \log p.$$

Solving the above inequalities for t we get,

$$t \geq \max\left\{\sqrt{\frac{4r\bar{\lambda}\tau \log p}{n}}, \frac{4r\tau \log p}{3n}\right\}.$$

So we can take,

$$t \geq \sqrt{\frac{4r\bar{\lambda}\tau \log p}{n}} + \frac{4r\tau \log p}{3n}.$$

Therefore,

$$\mathbb{P}\left(r\bar{\lambda} - \bar{y} < \sqrt{\frac{4r\bar{\lambda}\tau \log p}{n}} + \frac{4r\tau \log p}{3n}\right) \geq 1 - p^{-\tau}.$$

Now, solving the inequality,

$$r\bar{\lambda} - \bar{y} < \sqrt{\frac{4r\bar{\lambda}\tau \log p}{n}} + \frac{4r\tau \log p}{3n},$$

we get,

$$r\bar{\lambda} < 2\bar{y} + \frac{4\tau \log p}{n} + \frac{8r\tau \log p}{3n}.$$

Hence, (4.4) holds.

Proof of Lemma 4.5.2. Let $X \sim \chi^2(k)$. Then, $\mathbb{E}(X) = k$. Since, χ^2 distribution is a special case

of gamma distribution with parameters $(\frac{k}{2}, 2)$, the pdf of X is given by,

$$f(x) = \frac{1}{\Gamma(\frac{k}{2})2^{\frac{k}{2}}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}},\ 0 < x < \infty.$$

Now,

$$\begin{aligned}
\mathbb{E}\exp(t(x-k)) &= \int_0^\infty \exp(t(x-k)) \frac{1}{\Gamma(\frac{k}{2})2^{\frac{k}{2}}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} dx \\
&= \frac{\exp(-tk)}{\Gamma(\frac{k}{2})2^{\frac{k}{2}}} \int_0^\infty x^{\frac{k}{2}-1} e^{-x(\frac{1}{2}-t)} dx \\
&= \frac{\exp(-tk)}{\Gamma(\frac{k}{2})2^{\frac{k}{2}}} \frac{\Gamma(\frac{k}{2})}{(\frac{1}{2}-t)^{\frac{k}{2}}} \\
&= \left(\frac{\exp(-t)}{\sqrt{1-2t}}\right)^k \\
&\leq \exp(2kt^2) \text{ for } |t| < \frac{1}{4}. \text{ (using Bartlett's 3rd lecture).}
\end{aligned}$$

Hence the proof.

Proof of Lemma 4.5.3. Since, $y_i \sim$ Sub-exp$(4c_i, 4)$. Therefore,

$$\ln M_{y_i - \mathbb{E}(y_i)}(t) \leq \frac{t^2(4c_i)}{2} \text{ for } |t| < \frac{1}{4}.$$

$$
\begin{aligned}
\ln M_{\sum_{i=1}^n a_i[y_i-\mathbb{E}(y_i)]}(t) &= \ln \mathbb{E}\left(e^{t\sum_{i=1}^n a_i[y_i-\mathbb{E}(y_i)]}\right) \\
&= \ln\left[\prod_{i=1}^n \mathbb{E}\left(e^{ta_i[y_i-\mathbb{E}(y_i)]}\right)\right] \\
&= \sum_{i=1}^n \ln \mathbb{E}\left(e^{ta_i[y_i-\mathbb{E}(y_i)]}\right) \\
&= \sum_{i=1}^n \ln M_{y_i-\mathbb{E}(y_i)}(a_i t) \\
&\leq \frac{\sum_{i=1}^n t^2\cdot(a_i^2\cdot 4c_i)}{2} \text{ for } |t| < \frac{1}{4\max_i |a_i|} \\
&\leq \frac{t^2(\max_i |a_i|)^2\sum_{i=1}^n 4c_i}{2} \text{ for } |t| < \frac{1}{4\max_i |a_i|} \\
&= \frac{t^2\cdot 4(\max_i |a_i|)^2\sum_{i=1}^n c_i}{2} \text{ for } |t| < \frac{1}{4\max_i |a_i|}.
\end{aligned}
$$

Hence,

$$
\sum_{i=1}^n a_i y_i \sim \text{ Sub-exp}\left(4(\max_i |a_i|)^2\sum_{i=1}^n c_i, 4\max_i |a_i|\right).
$$

Proof of Lemma 4.5.4. Again, since $y_i \sim$ Sub-exp $(4,4)$ for $i = 1, 2, \cdots, n$ we have,

$$
\mathbb{P}\left(\sum_{i=1}^n(\lambda_i - y_i) > t\right) \leq \begin{cases} \exp\left(-\frac{t^2}{8}\right), & \text{if } 0 \leq t \leq 1 \\ \exp\left(-\frac{t}{8}\right), & \text{if } t > 1. \end{cases}
$$

which gives,

$$
\mathbb{P}\left(\bar{\lambda} - \bar{y} > t\right) \leq \begin{cases} \exp\left(-\frac{n^2t^2}{8}\right), & \text{if } 0 \leq t \leq 1 \\ \exp\left(-\frac{nt}{8}\right), & \text{if } t > 1. \end{cases}
$$

Therefore,

$$\mathbb{P}\left(\bar{\lambda} - \bar{y} > t\right) \leq p^{-\tau},$$

if and only if,

$$\frac{n^2 t^2}{8} \geq \tau \log p, \text{ and, } \frac{nt}{8} \geq \tau \log p.$$

Solving both the above inequalities for t we get,

$$t \geq \max\left(\frac{8\tau \log p}{n}, \frac{\sqrt{8\tau \log p}}{n}\right).$$

Hence, we can consider,

$$t \geq \frac{8\tau \log p}{n} + \frac{\sqrt{8\tau \log p}}{n},$$

Therefore, (4.7) holds.

CHAPTER 5: CONCLUSION AND FUTURE WORK

In this dissertation, we justified the application of lasso to the solution of ill-posed linear inverse problems. In particular, we consider application of the lasso with random dictionaries and Gaussian noise. We also studied the application of lasso when the error distribution is not Gaussian. We provided theoretical guarantees and evaluated the technique via a limited simulation study. While in the first part (Gaussian error) our oracle and cross-validation provided estimators with better precision than the oracle SVD for a low sample size, this not true for the case of non-Gaussian errors. We hope to improve our precision for non-Gaussian errors with a low sample size as a part of future work.

LIST OF REFERENCES

[1] Abramovich, F., Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems.

[2] Abramovich, F., Pensky, M., Rozenholc, Y. (2013). Laplace deconvolution with noisy observations. *Electronic Journal of Statistics*, **7**, 1094-1128

[3] Bickel, P.J., Ritov, Y., Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **3**7, 1705 - 1732.

[4] Brown, L. D., Cai, T. T., Zhou, H. H. (2010). Non parametric regression in exponential families.

[5] Bühlmann, P., van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer.

[6] Bunea, F., Tsybakov, A., Wegkamp, M. (2007) Sparsity oracle inequalities for the Lasso.

[7] Candès, E. J., Eldar, Y., Needell, D., Randall, P. (2010). Compressed sensing with coherent and redundant dictionaries. *Appl. Computat. Harmonic Anal.*, **31**, 59–73.

[8] Cavalier, L., Golubev, G.K., Picard, D., Tsybakov, A.B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, **30**, 843-874.

[9] Cavalier, L., Golubev, Yu. (2006) Risk hull method and regularization by projections of ill-posed inverse problems.

[10] Cavalier, L., Reiss, M. (2014). Sparse model selection under heterogeneous noise: Exact penalisation and data-driven thresholding.

[11] Cohen, A., Hoffmann, M., Reiss, M. (2004). Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM Journ. Numer. Anal.*, **42**, 1479–1501.

[12] Comte, F., Cuenod, C. -A., Pensky, M., Rozenholc, Y. (2017). Laplace deconvolution on the basis of time domain data and its application to Dynamic Contrast Enhanced imaging. *Journ. Royal Stat. Soc., Ser.B*, **79**, 69–94.

[13] Donoho, D.L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition.

[14] Dalalyan, A.S., Hebiri, M., Lederer, J. (2014). On the prediction performance of the Lasso. ArXiv: 1402.1700

[15] Efromovich, S., Koltchinskii, V. (2001). On inverse problems with unknown operators.

[16] Foucart, S., Rauhut, H. (2013). A Mathematical Introduction to Compressive Sensing. Springer, New York.

[17] Fryzlewicz, P., Nason, G. P. (2004) A Haar-Fisz Algorithm for Poisson Intensity Estimation.

[18] Gockenbach, M. (2016). *Linear Inverse Problems and Tikhonov Regularization.* The Mathematical Association of America.

[19] Golubev, Y. (2010). On universal oracle inequalities related to high-dimensional linear models.

[20] Gupta, P., Pensky, M. (2018). Solution of linear ill-posed problems using random dictionaries.

[21] Hoffmann, M., Reiss, M. (2008) Nonlinear estimation for linear inverse problems with error in the operator.

[22] Kalifa, J., Mallat, S. (2003). Thresholding estimators for linear inverse problems and deconvolutions.

[23] Kolaczyk, E. D. (1999) Bayesian Multiscale Models for Poisson Processes.

[24] Kolaczyk, E. D., Nowak, R. D. (2004) Multiscale likelihood analysis and complexity penalized estimation.

[25] Kroll, M. (2016). Concentration inequalities for Poisson point processes with application to adaptive intensity estimation.

[26] Le Pennec, E., Mallat S. (2005). Sparse geometric image representations with bandelets.

[27] Mairal, J. (2014). *SPAMS: a Sparse Modeling Software*, MatLab toolbox. `http://spams-devel.gforge.inria.fr`

[28] Pensky, M. (2016). Solution of linear ill-posed problems using overcomplete dictionaries.

[29] Reynaud-bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities.

[30] Tropp, J.A., Wright, S. J. (2010). Computational methods for sparse solution of linear inverse problems. **98**, 948-958.

[31] Tsybakov, A. B.(2009). Introduction to Nonparametric Estimation.

[32] Vareschi T. (2013). Noisy Laplace deconvolution with error in the operator.

[33] Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices.* Cambridge University Press.

[34] Willett, R. M., Raginsky, M. (2009) Performance Bounds on Compressed Sensing with Poisson Noise.