High Dimensional Statistical Models: Applications to Climate

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Soumyadeep Chatterjee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ARINDAM BANERJEE

September, 2015

Acknowledgements

I would like to express my sincerest gratitude to my advisor, Prof. Arindam Banerjee, for his support and guidance throughout my graduate study. I will be forever thankful to him for introducing me to machine learning, and being a role model for doing research. His enthusiasm for learning is infectious, and I have learnt from him in more ways than I can possibly say.

I would like to thank Prof. Vipin Kumar, Prof. Snigdhansu Chatterjee and Prof. Daniel Boley, for being my dissertation committee members. I would also like to extend my gratitude to my collaborators Auroop Ganguly, Debasish Das and Karsten Steinhaeuser.

Life as a grad student would not have been the same without my great labmates. My heartfelt thanks to Qiang, Huahua, Puja, Rudy, Hanhuai, Amir, Vidyashankar, Farideh, Konstantina, Sheng, Igor, Nick, Andre and Karthik. I enjoyed all the discussions in the lab and corridors of the department!

Warm thanks to my incredible friends who made Minneapolis more beautiful in summer, and warmer in winter! Thanks to Ayan, Sauptik, Somnath, Koustav, Sanjoy and of course, the great Mr. Shameek Bose and Mr. Kaushik Basu!

I am who I am because of my parents. They sacrificed more than I could ever measure, and have been my support for all my life. Although a thank you will not do justice, I am grateful for having them, and will forever be in their debt.

I am thankful for having the best sister in the world, and for all the silly little things that we laugh about. Life would not be the same without our constant quibbling and silly jokes!

For Kasturi, a special thanks for being my best friend through all these years, for rejoicing in my successes more than me, and for being my support all through. I have grown to be a better person because of her.

Dedication

To my father Dr. Manoranjan Chatterjee, and my mother, Dr. Shibani Chatterjee

Abstract

Recent years have seen enormous growth in collection and curation of datasets in various domains which often involve thousands or even millions of variables. Examples include social networking websites, geophysical sensor networks, cancer genomics, climate science, and many more. In many applications, it is of prime interest to understand the dependencies between variables, such that predictive models may be designed from knowledge of such dependencies. However, traditional statistical methods, such as least squares regression, are often inapplicable for such tasks, since the available sample size is much smaller than problem dimensionality. Therefore we require new models and methods for statistical data analysis which provide provable estimation guarantees even in such high dimensional scenarios. Further, we also require that such models provide efficient implementation and optimization routines. Statistical models which satisfy both these criteria will be important for solving prediction problems in many scientific domains.

High dimensional statistical models have attracted interest from both the theoretical and applied machine learning communities in recent years. Of particular interest are parametric models, which considers estimation of coefficient vectors in the scenario where sample size is much smaller than the dimensionality of the problem. Although most existing work focuses on analyzing sparse regression methods using L1 norm regularizers, there exist other "structured" norm regularizers that encode more interesting structure in the sparsity induced on the estimated regression coefficients. In the first part of this thesis, we conduct a theoretical study of such structured regression methods. First, we prove statistical consistency of regression with hierarchical tree-structured norm regularizer known as hiLasso. Second, we formulate a generalization of the popular Dantzig Selector for sparse linear regression to any norm regularizer, called Generalized Dantzig Selector, and provide statistical consistency guarantees of estimation. Further, we provide the first known results on non-asymptotic rates of consistency for the recently proposed k-support norm regularizer. Finally, we show that in the presence of measurement errors in covariates, the tools we use for proving consistency in the noiseless setting are inadequate in proving statistical consistency.

In the second part of the thesis, we consider application of regularized regression methods to statistical modeling problems in climate science. First, we consider application of Sparse

Group Lasso, a special case of hiLasso, for predictive modeling of land climate variables from measurements of atmospheric variables over oceans. Extensive experiments illustrate that structured sparse regression provides both better performance and more interpretable models than unregularized regression and even unstructured sparse regression methods. Second, we consider application of regularized regression methods for discovering stable factors for predictive modeling in climate. Specifically, we consider the problem of determining dominant factors influencing winter precipitation over the Great Lakes Region of the US. Using a sparse linear regression method, followed by random permutation tests, we mine stable sets of predictive features from a pool of possible predictors. Some of the stable factors discovered through this process are shown to relate to known physical processes influencing precipitation over Great Lakes.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent years have seen vast increase in the rate of data generation and storage in various fields of science and technology. For example, deployment of new sensors and satellite have had a major impact on high quality atmospheric and geological data collection. The multitude of rapidly expanding social network sites store petabytes of user generated content in servers across the globe. Owing to the massive size of data available, high dimensional statistical models are increasingly being used in various machine learning and data mining tasks. Often, such datasets involve measurements of certain variables over space and time. and there is growing interest in understanding statistical dependencies between such variables. Moreover, many domains require modeling multi-resolution and correlated variables, and such inherent structure in the data needs consideration when designing and learning the models. Modern machine learning approaches for high dimensional modeling have successfully been applied to multiple such domains, such as signal processing [3, 4, 5], bioinformatics [6, 7], computational biology [8, 9], astronomy [10], web data analysis [11, 12], querying [13, 14], and many more. Further, new applications of these models are increasingly finding novel applications in scientific domains, such as climate science [15, 16], brain imaging [17, 18], sensor networks [19, 20], where classical analysis techniques are being challenged when dealing with high dimensional data and the generative complex phenomena .

1.1 Modeling in high dimensions with few samples

The central goal of statistical modeling is to capture the *dependencies* or "relatedness" between variables. For example, consider the problem of understanding the variation of precipitation over North America with the variations in sea surface temperature (SST) over the tropical Pacific Ocean [21]. We want to understand, first, whether there exists any dependence between these two quantities. Second, if there exists any, we want to quantify the degree to which these are dependent. Now, there are a number of climate processes and variables which influence precipitation over the U.S., and influence over precipitation is a combined effect of all such variables. In general, we often encounter such *regression* problems, where the goal is to model a response variable y with multiple predictors or covariates $x_1, \ldots, x_p$.

The central challenge in high dimensional modeling arises from the small sample size n relative to the dimensionality p of the problem. For example, in a high dimensional regression problem, often the set of possible predictors can run into thousands or even millions [22], while the number of samples available to train a model is much smaller, often by orders of magnitude. Consider the Ordinary Least Squares (OLS) estimator. Given samples $(\mathbf{y}, \mathbf{X})$ from the response and the covariates, OLS estimates the following regression coefficient vector

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{1.1}$$

However, when $n < p$, the estimation problem is ill-posed, and OLS estimation does not work. In general, problem arises when one tries to analyze traditional statistical estimators in the context of high dimensional models. Often, the traditional estimators require that the model be *fully determined*, i.e. $n \gg p$, and provide asymptotic bounds on estimation error when $n \to \infty$. Both these assumptions fail in the high dimensional regime, and given the small sample size n.

Scientific domains, such as climate science, are often plagued with such ill-posed estimation problems. For example, the number of possible covariates for predicting precipitation is huge. However, the availability of high quality climate data is limited to only the past few decades. Therefore, high dimensional statistical models are the key to understanding and modeling the dependencies between different climate variables. Further, we also require that the estimates be *stable*, so that we can draw interpretable physical hypotheses from the statistical estimates. Moreover, we often understand qualitative structures about the model. For example, it is well known that convective precipitation is dependent on the land surface temperature and moisture

availability at a location. Such constraints should inform the estimator of possible dependencies. The motivation of this work lies in developing analyzing methods to solve these problems.

1.2 A Brief History

Modern machine learning approaches aim to solve the ill-posed estimation problems by imposing structural constraints on the parameter to be estimated. One of the earliest examples is in high dimensional regression, where it was suggested to constrain or bound the L_2 norm of the regression parameter . It is known as the *ridge regression* problem, and is widely used for solving under-determined regression problems [23, 24]. However, although constraining the L_2 norm as in ridge regression enables us to solve a high dimensional regression problem, it *does not* impose any control on model complexity since all features in $\mathbf{X}$ are typically assigned non-zero weights. One way of controlling model complexity is to restrict the number of coefficients that are assigned non-zero weights, viz. doing *feature-selection* while fitting the regression model. This can be achieved by constraining the L_0 norm of the regression coefficient as $\|\boldsymbol{\beta}\|_0 \leq k$, where $\|\boldsymbol{\beta}\|_0$ denotes the number of non-zero elements in $\boldsymbol{\beta}$.

The formulation is closely related to the *basis pursuit* problem in signal processing where one wants to reconstruct a sparse signal from a few measurements, given a fixed sampling matrix $\mathbf{X}$. The above sparse regression problem, though, suffers from a serious drawback. It is non-convex, and for large p, practically impossible to solve directly. Much effort has been spent to design approximations of this problem, and to show that such approximations yield comparable results. The major breakthrough in machine learning and statistics has been to show that under some mild conditions, the sparse regression problem can be exactly and efficiently solved by considering the L_1 norm to be a relaxation of the L_0 norm and therefore constraining the L_1 norm of the estimate. It is known as the Least Absolute Shrinkage and Selection Operator, or commonly, LASSO [25, 26, 27]. The LASSO method has found great success in applications to various domains [26]. In particular, sparsity as a structural constraint is now extensively used for feature selection to control model complexity.

1.3 Theoretical Advances

Scientific fields, such as climate science, often deal with complex phenomena, where simple structural constraints, such as imposed by L_1 norm, are inadequate, and/or require additional procedures in order to ensure stable parameter estimation. For example, if there exist certain groups of covariates $G \in \mathcal{G},\ G \subseteq \{1, 2, \ldots, p\}$ which *jointly* activate and influence the response y, L_1 norm regularization may lead to *unstructured* sparsity, so that only certain members of a group are assigned non-zero coefficients. Therefore, recent work has concentrated on developing regularizers which respect such structure in the data. For example, the group-Lasso regularizer [28, 29] considers a mixed L_1/L_2 penalty

$$\|\boldsymbol{\beta}\|_{(1,\mathcal{G})} = \sum_{G \in \mathcal{G}} \|\boldsymbol{\beta}_G\|_2 \ , \tag{1.2}$$

which considers penalization over groups of coefficients rather than singletons, so that feature selection occurs over groups and the resulting model is more interpretable and often has better prediction performance [29]. More generally, one may consider different *norm regularizers* $\mathcal{R}(\boldsymbol{\beta})$ over the regression coefficient vector $\boldsymbol{\beta}$ where $\mathcal{R}$ encodes an appropriate structure. Such structures may be qualitatively known from domain knowledge (e.g. groups over spatial locations considered in Chapter6) or even latent structures over coefficients (e.g., the k-support norm in Chapter 4).

For any such regularized regression problem, we are often interested in the "quality of estimation", i.e. assuming a generative model for the data, whether the estimated $\hat{\boldsymbol{\beta}}$ is "close" to the statistical parameter. For example, often assumed is a linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \ , \tag{1.3}$$

where $\boldsymbol{\beta}^*$ is the statistical parameter, and $\boldsymbol{\epsilon}$ is an additive random noise variable. Under this model, we aim to upper bound the error in estimation $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$. For the high dimensional regime, of particular interest are *non-asymptotic* upper bounds on the estimation error. Such bounds hold for finite n and p, and decrease with a certain rate $F(n, p)$ with high probability. The rate $F(n, p)$ is called the "rate of consistency". For example, if $\boldsymbol{\beta}^*$ is an s-sparse vector and we solve the LASSO problem to obtain $\hat{\boldsymbol{\beta}}$, we obtain the following bound with high probability [30, 31]

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 = \mathrm{O}\left(\sqrt{\frac{s \log p}{n}}\right) \ . \tag{1.4}$$

The limitation of existing methods lies in the type of structures that can be modeled in the estimation problem. For most of the research conducted in this domain, the focus was on sparsity, or group-sparsity. However, there exists more involved structures, such as hierarchies, which cannot be modeled by sparse or group-sparse regularizers. Some recent developments suggest using structured regularizers, such as hiLasso [32, 33], or k-support norm to alleviate this problem. However, no theory of consistency exists for such structured regularizers. Further, it is unknown whether the existing statistical tools are adequate for analyzing such estimators. For certain estimators, such as the Dantzig Selector [34, 35], it is not even known how to modify the estimators to incorporate general structured norm regularizers $\mathcal{R}$.

1.4 Applications to Climate Science

A core research question in climate science is improved understanding of interactions of geophysical, atmospheric and ocean variables in order to more accurately simulate future climate system of the earth. To this effect, General Circulation Models (GCM) [36] have been developed in the climate science community to simulate future climate, and obtain an overview of possible climate scenarios. However, the output resolutions of these models are typically too coarse in scale to be useful in practical comprehensive planning situations, such as applying hydrological modeling in flood-risk analysis [37, 38]. One possible solution is to downscale the output of the GCMs to a higher resolution in space/time, and using the high-resolution downscaled data for use in climate-change impact assessment studies. Traditionally, downscaling has been done by constructing higher resolution physical dynamical models, and executing simulations with boundary conditions forced by either the GCMs or observed measurements [39, 40]. Such dynamical downscaling suffers from the limitation that each simulation run is computationally expensive even for relatively small regions on the globe. Therefore, statistical downscaling, which tries to find statistical relationships between coarse resolution GCM outputs and high resolution predictands is increasingly becoming popular for climate downscaling [41, 42, 43].

A variable of considerable interest is precipitation, which is known to be difficult to simulate in GCMs [44, 45]. Therefore, statistical downscaling of precipitation is an area of active research in the climate science community [41, 43]. The key requirement in statistical downscaling of precipitation is improved understanding of dependencies between local precipitation large scale climate variables and processes. Regression methods, which consider large scale

climate variables are predictors, and the local variable as predictand, often face the problem of high dimensionality, because of the enormous number of possible predictor variables, and limited high quality observational data available from the past thirty years [33]. Although various machine learning methods have previously been used in climate science for improved prediction performance [46, 24], there are two key issues the belie the gain in predictive performance. First, the models often require extensive data pre-processing and manual selection of covariates [47, 48, 49] in order to show improvements. Second, training complex models on small datasets often leads to overfitting, and therefore lack of interpretability. Since a key goal is development of physical hypothesis on mechanisms affecting climate phenomenon, both these issues demand attention for advancing statistical modeling in climate science.

1.5 Roadmap of the thesis

In this thesis, we try to address the theoretical limitations discussed in Section 1.3, and develop method for solving statistical modeling problems in climate science, as discussed in Section 1.4. We start by reviewing the state of the art and related work in theory and applications in Chapter 2. The rest of the thesis is divided into two parts: (1) Theoretical Developments and (2) Applications.

In Part I we study structured regression methods, and associated non-asymptotic statistical consistency guarantees. In chapter 3, we consider hierarchical tree-norms which generalize the sparse and group-sparse norms. The estimator, called the hierarchical Lasso (hiLasso), encodes a tree-structure hierarchy in the sparsity induced on the regression coefficient $\boldsymbol{\beta}$. Using some recently developed analysis tools in statistics and machine learning, we prove rates of convergence for estimators regularized by the hiLasso regularizer. The *Sparse Group Lasso* is a special case of hiLasso, and empirical results have shown that it performs better than LASSO in a number of application domains. We prove statistical consistency for the hiLasso and the Sparse Group Lasso, and discuss existing methods for efficient optimization.

The Dantzig Selector is an estimator for sparse linear regression problems, and was formulated as an alternative to LASSO. Particularly, it involves solving the following linear program

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \ \|\boldsymbol{\beta}\|_1 \tag{1.5}$$

$$\text{s. t.} \quad \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \lambda_n \ . \tag{1.6}$$

Interestingly, the analysis of the Dantzig Selector shows close similarities to LASSO in the rates of consistency obtained on the estimation error $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$ [35]. In Chapter 4, we consider a generalization of the Dantzig Selector by considering *any* norm regularization $\mathcal{R}(\boldsymbol{\beta})$ instead of the L_1 norm. Particularly, we show that the Generalized Dantzig Selector (GDS) [50] is statistically consistent, and we derive non-asymptiotic rates of convergence for some well-known structured regularizers, such as the group-lasso norm. Further, we consider the recently proposed k-support norm, which considers sparsity over latent groups of coefficients. This norm is of particular interest, since empirical results have shown that predictive performance improves with the k-support norm estimator over LASSO, in scenarios with correlated covariates. We prove sharp rates of consistency for the k-support norm, and also provide numerical simulations to illustrate its properties.

Statistical analysis of regularized estimators require that samples $\mathbf{X}, \mathbf{y}$ are not corrupted by noise. However, in real world, one often encounters scenarios where such assumptions fail. For example, miscalibration of sensors in a sensor network may lead to incorrect, noisy or biased measurements. It is unknown whether the statistical analysis of estimators still provides consistency guarantees under such noisy scenarios. In Chapter 5, we consider the scenario when the predictors $\mathbf{X}$ may themselves be corrupted with additive noise. Particularly, we show that in the noisy setting, the current analysis tools fail to provide consistency, and the error bound obtained does not decrease to zero with increasing sample size. However, we also show that if an estimate of the noise covariance is available, then with an appropriate correction, GDS provides consistent estimates.

In Part II of the thesis, we consider applications of structured regression methods to statistical modeling in climate science. The goal is to use such statistical models, trained on observation data, to extend our understanding of predictability in climate, and providing insights for improving the physical models whose simulation outputs are the only available data for future climate scenarios. As the first application, in Chapter 6, we study the interaction between climate variables on land and over oceans. We model the problem as a regression problem where each variable over land is a response, while multiple gridded variables over oceans are predictors. The resulting model is high dimensional, and therefore lends itself to the structured regularizers we develop in earlier chapters. Particularly, we consider the Sparse Group Lasso regularizer and consider groups of predictors over each geographical grid point on oceans. Extensive experiments are conducted on data obtained from the NCEP Reanalysis project [51],

and we illustrate that adding structured regularizers not only improves model interpretability, but also provides better predictive performance than both unregularized models and unstructured regularizers.

The second study considers an important question in applying statistical modeling in climate science: how can statistical modeling be used to generate physical hypotheses on dependencies between climate processes? As a use case, we consider winter precipitation over the Great Lakes region in the US, with the goal of finding the dominant factors for precipitation from among local, regional and global climate processes. In order to identify the dominant factors, we utilize a random permutation test to obtain "stable" sets of features selected from a large pool of possible predictors. Extensive experiments and analysis carried out using data obtained from weather stations, reanalysis products [51] and derived climate indices illustrate two central observations. First, regularized statistical models, aided by stability tests, are powerful tools which can mine dominant factors for complex phenomenon such as precipitation. For example, 700hPa pressure level differences over the north pacific and north america in autumn were discovered to be important drivers of winter precipitation over the Eastern Great Lakes. Further, winter minimum temperature over the Great Lakes was also determined to be a dominant factor for variations in snowfall. These discovered factors can be further studied as possible hypotheses from the point of view of climate science for further verification and possible insights into climate processes that affecting precipitation. Second, and more importantly, rich representative features are the most important component of designing useful statistical models. For example, in order to capture the effect of an ocean oscillation such as El-Nino [51], it is of utmost importance to include indices that are rich representations of this phenomenon. If the features do not adequately represent the phenomenon, no statistical model may be powerful enough to capture the influence. Finally, we end with conclusions in Chapter 8.

Chapter 2

Related Work

2.1 Theory of High Dimensional Models

One of the earliest research to devote attention to solving ill-posed regression problems can be attributed to Tikhonov [52], who proposed using the L_2 norm regularization. It has since then been called ridge regression [23] and Tikhonov regularization [53]. Specifically, given a response vector $\mathbf{y} \in \mathbb{R}^n$ and a matrix of covariates $\mathbf{X} \in \mathbb{R}^{n \times p}$, we solve the following regression problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{2.1}$$

$$\text{s. t. } \|\boldsymbol{\beta}\|_2 \leq \lambda \,, \tag{2.2}$$

where $\boldsymbol{\beta}$ is a regression parameter to be estimated, and $\lambda > 0$ is a *regularization* parameter that can be chosen by the user. This particular problem and its equivalent counterpart

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \,, \tag{2.3}$$

are known as the ridge regression problem.

However, although constraining the L_2 norm as in ridge regression enables us to solve a high dimensional regression problem, it *does not* impose any control on model complexity since all features in $\mathbf{X}$ are typically assigned non-zero weights. One way of controlling model complexity is to restrict the number of coefficients that are assigned non-zero weights, viz. doing *feature-selection* while fitting the regression model. This can be achieved by constraining

the L_0 norm of the regression coefficient

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \ \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{2.4}$$

$$\text{s. t. } \|\boldsymbol{\beta}\|_2 \leq 1, \qquad \|\boldsymbol{\beta}\|_0 \leq k \ , \tag{2.5}$$

where $\|\boldsymbol{\beta}\|_0$ denotes the number of non-zero elements in $\boldsymbol{\beta}$.

The above sparse regression problem, though, suffers from a serious drawback. It is non-convex, and for large p, practically impossible to solve directly. Much effort has been spent to design approximations of this problem, and to show that such approximations yield comparable results. The major breakthrough in machine learning and statistics has been to show that under some mild conditions, the sparse regression problem can be exactly and efficiently solved by considering the L_1 norm to be a relaxation of the L_0 norm and solving the following convex relaxation:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \ \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \tag{2.6}$$

$$\text{s. t. } \|\boldsymbol{\beta}\|_2 \leq 1, \qquad \|\boldsymbol{\beta}\|_1 \leq \lambda_n \ , \tag{2.7}$$

where $\lambda_n > 0$ is a regularization parameter. The equivalent unconstrained problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \ \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_n\|\boldsymbol{\beta}\|_1 \tag{2.8}$$

is known as the Least Absolute Shrinkage and Selection Operator, or commonly, LASSO [27, 25]. However, the idea of feature selection, where one selects a subset of the features in a regression problem, is not new, but has received attention from the machine learning community [54, 55, 56, 57]. However, most of these methods suffered from non-convexity and lack of efficient optimization algorithms, and often provided sub-optimal solutions.

In a parallel development, the *basis pursuit* algorithm was developed in the context of compressed sensing to recover sparse signals in under-determined systems [3, 58]. The basis pursuit problem proposes solving the following optimization problem:

$$\min_{\boldsymbol{\beta}} \ \|\boldsymbol{\beta}\|_1 \tag{2.9}$$

$$\text{s.t. } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} \ . \tag{2.10}$$

This problem may be understood as solving the sparse regression problem in the absence of the additive noise $\boldsymbol{\epsilon} = 0$. It was shown that under certain conditions, called the *restricted isometry*

property (RIP), of the matrix $\mathbf{X}$, basis pursuit can exactly recover the s-sparse parameter $\boldsymbol{\beta}^*$ which generated the data $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^*$ [58]. Specifically, the RIP property requires that the singular values of the all sub-matrices $\mathbf{X}_S$ of matrix $\mathbf{X}$ with s columns be bounded as

$$(1 - \delta_S)\|\boldsymbol{\beta}_S\|_2^2 < \|\mathbf{X}_S\boldsymbol{\beta}_S\|_2^2 < (1 + \delta_S)\|\boldsymbol{\beta}_S\|_2^2 \,, \tag{2.11}$$

for all vectors $\boldsymbol{\beta}_S \in \mathbb{R}^s$ and for some $\delta_S > 0$. However, the compressed sensing community did not consider random design matrices $\mathbf{X}$.

No proof of consistency existed for sparse regression with random design matrices till the first proof of the statistical consistency of LASSO was provided in [59]. Although [60] had shown that LASSO is consistent in estimating the support of $\boldsymbol{\beta}^*$, consistency with respect to the error $\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2$ was not yet proved. [59] proved that under assumptions of Gaussianity of the covariate or design matrix $\mathbf{X}$, and the noise $\boldsymbol{\epsilon}$, LASSO is consistent.

In the following years, there has been increasing interest in designing regularizers that encode more complex structures than simple sparsity. For example, [28] proposed the group-lasso regularizer, which considers sparsity over groups of variables. Further, they proved statistical consistency of regularized estimators with the group-lasso norm. A number of papers in the following years extended this work to groups with overlaps and proposed efficient methods for optimization [29, 61, 62, 63]. The hierarchical group Lasso, and its special case, the Sparse Group Lasso, were proposed in [64, 32]. Such hierarchical norms encoded a tree-structured hierarchy in the sparsity induced in the regression estimate. Further, fast optimization methods were developed in [65, 66] for efficient proximal projections with the hierarchical lasso norm. However, although much progress was made in terms of developing novel sparse regularizers and associated optimization methods, no consistency results existed in the literature for such regularizers.

In a parallel development in the statistics community, the Dantzig Selector was proposed as a non maximum likelihood estimator for sparse linear regression [34]. In contrast to LASSO, which involves an unconstrained optimization problem, the Dantzig Selector is similar in spirit to the basis pursuit method [58] and involves the following estimation problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \ \|\boldsymbol{\beta}\|_1 \tag{2.12}$$

$$\text{s.t.} \ \ \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \lambda_n \,. \tag{2.13}$$

Similar to LASSO, the Dantzig Selector has been extended to group-sparse norms [67]. Further, [35] has analyzed and compared the Dantzig Selector with LASSO to show that for sparse regression problems, they behave similarly, and one obtains similar statistical consistency guarantees. However, the Dantzig Selector has not been extended to other norm regularizers, and in Chapter 4 we provide the first results on generalizing Dantzig Selector to *any* norm $\mathcal{R}$.

Although multiple papers were written proving the statistical consistency of LASSO, group-lasso etc. there did not exist a unified framework for proving statistical consistency of sparse regression methods. [30] provided the first unified framework for analyzing sparse, or in general, decomposable regularizers. A decomposable norm regularizer $\mathcal{R}$ is such that if M is a subspace in $\mathbb{R}^p$, and $M^\perp$ is its orthogonal space, then for any vector $\boldsymbol{\beta} = \boldsymbol{\beta}_M + \boldsymbol{\beta}_{M^\perp}$, $\mathcal{R}(\boldsymbol{\beta}) = \mathcal{R}(\boldsymbol{\beta}_M) + \mathcal{R}(\boldsymbol{\beta}_{M^\perp})$. The paper presented two key conditions required for proving consistency of regularized estimators. The first condition was the notion of the error cone, which shows that for an appropriate choice of the regularization parameter λ_n, the error vector $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ lies in a cone $\mathbb{C}_{\mathcal{R}}$, whose shape depends on the regularizer $\mathcal{R}$. The second key condition is known as Restricted Eigenvalue (RE) condition, which requires the design matrix $\mathbf{X}$ to satisfy the lower bound

$$\|\mathbf{X}\Delta\|_2 \geq \kappa \|\Delta\|_2 \ , \ \forall \Delta \in \mathbb{C} \tag{2.14}$$

such that $\kappa > 0$. This condition is less strict than the RIP condition described before, but is essential in determining the sample complexity for consistent estimation. Their analysis borrowed results from [68], which showed that Gaussian random matrices satisfy this condition with high probability. However, the analysis technique did not extend to non-decomposable norms, such as the overlapping group lasso norm.

The framework for analyzing general norm regularizers was put forward in [69], albeit in a different form. [69] considered a class of norms known as *atomic norms*, which have previously been proposed in the context of basis pursuit [58]. For a set of atoms $\mathcal{A} = \{\mathbf{a}\}$, $\mathbf{a} \in \mathbb{R}^p$, the atomic norm is defined as

$$\|\boldsymbol{\beta}\|_{\mathcal{A}} = \inf \left\{ \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \ : \ c_{\mathbf{a}} \geq 0, \ \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a} = \boldsymbol{\beta} \right\} . \tag{2.15}$$

[69] proposed the following minimization problem to recover a statistical parameter $\boldsymbol{\beta}^*$ supported on a *small* number of atoms $\mathcal{A}^* \subseteq \mathcal{A}$:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \ \|\boldsymbol{\beta}\|_{\mathcal{A}} \tag{2.16}$$

$$\text{s. t.} \ \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 \leq \lambda_n \ . \tag{2.17}$$

The key idea developed in this paper was to analyze the sample complexity of estimation using the *Gaussian width* of a set [70, 71, 72, 73]. The Gaussian width provides a measure of complexity of sets, particularly norm balls of the regularizer $\mathcal{R}$. [71, 72, 73] has shown that the Gaussian width also arises in showing the RE condition for Gaussian design matrices. [69] provided some new results on bounding the Gaussian width with some analytical quantities dependent on the dimensionality p, and the sample size n.

2.2 Optimization Methods

Optimization methods for regularized regression has received much attention from the machine learning community in recent years. The LARS algorithm for LASSO [25] was seminal in that it provided understanding of the geometry of LASSO solutions. Development of efficient methods started with increasing interest in proximal gradient methods [74, 75, 76, 77]. LASSO, for example, provided fast algorithms through iterative shrinkage [78], since the lasso proximal operator can be computed in closed form [79]. The proximal method has been shown to be efficient for many other norm regularized regression problems. For example, [76] developed an efficient dual ascent algorithm for hierarchical sparse norms based on an efficient proximal operator of the dual norm. On a similar note, [65] developed a fast primal algorithm based on proximal updates for hierarchical and group-structured norm regularizers, with efficient projections on the hierarchical tree encoded by the regularizer. In context of online learning [80, 81], forward-backward splitting algorithms are based on proximal methods, and provide fast stochastic optimization methods for large scale problems. Further, alternating direction method of multipliers (ADMM) [82, 83] provide an alternative to splitting methods for efficient optimization.

The Orthogonal Matching pursuit algorithm [4] was developed as a greedy algorithm for solving the noisy basis pursuit problem. The method greedily selects each dimension (or atom)

which minimizes the residual of a linear regression problem while being orthogonal to previously selected dimensions (or atoms). In the same vein the greedy forward-backward algorithm [84] aims to solve the L_1 regularized least squares problem by successively adding and removing features from a feature set, such that the residual norm is minimized.

More recently, coordinate descent [85] methods are receiving attention from the community. Although coordinate descent methods for lasso have previously been developed [86], recent research illustrates the ability to parallelize coordinate descent methods for sparse regression [87]. Further, randomized block coordinate methods [88] offer scope for massive parallelism for solving regularized regression problems. Taken together, such methods hold promise for efficient implementation in large scale problems.

2.3 Applications of Machine Learning models to climate

In recent years statistical modeling is receiving attention from the climate science community for improving predictive performance of traditional physical models [89], as well as for statistical downscaling [90]. Ridge regression, particularly, has been widely used for multi-model ensemble forecasting with General Circulation Models (GCM) [46, 24, 89], and for modeling transformation functions for computing surface temperature from satellite data [91]. However, regression with dimensionality reduction or feature selection has often been used in the context of statistical downscaling [47]. Most commonly, regression methodologies involve application of principal component analysis (PCA) to covariates to reduce dimensionality, followed by multivariate linear or non-parametric regression models on the principal component scores [47, 48, 49, 92, 15]. Feature selection, however, has received less attention in the community, and few papers exist on this topic [16].

Part I

Theoretical Developments

Chapter 3

Hierarchical Sparse Models

3.1 Introduction

The success of data mining techniques in complementing and supplementing findings from several topics of scientific research is well documented [6, 93, 94]. Often, the problems of interest involve learning relationships among the response and predictor variables, where the set of predictor variables may be very large. Recent work has proved the utility of having parsimony in the inferred dependency structure. Efforts in this direction have been successful in developing *sparse* models, which promote sparsity within the dependencies characterized by the model. These models have been applied successfully in a number of fields, such as signal processing [95], bioinformatics [96], computer vision [97] etc. Incorporating sparsity within a statistical model provides a natural control over the complexity of the model.

The classical statistical model trains from the training data at hand by defining a loss function to measure the discrepancy between its predictions and observations of the response variables. Optimization routines are used to obtain an optimal parameter set for the model so that the loss function is minimized. *Sparsity* is induced within the optimal parameter set by adding a *sparsity-inducing* regularizer function to the loss and optimizing this combination over the parameter set. The regularizer is usually a norm function of the parameter vector. This construction gives rise to a family of *sparse statistical models* with a convex loss function and a convex norm regularizer [26, 27]. Building on this literature, recent work has shown the utility of imposing *structure* among the dependencies through the use of group [28] and hierarchical norm regularizers [98, 64]. These structures can be learnt from some external sources, such

as domain experts, and are useful in obtaining more robust and interpretable predictive models. Efficient optimization algorithms have been proposed to solve such estimation problems [65]. Recent results [30, 59] have proved statistical consistency guarantees for a class of sparse estimators under fairly mild conditions.

In this chapter, we consider structured sparse estimator called the hierarchical Lasso (hiLasso), which encodes a tree-structured hierarchy in the induced sparsity. Using the analysis method developed in [30], we prove statistical consistency guarantees for the class of tree-structured hierarchical norm regularized estimation problems [65]. The chapter is arranged as follows. We formally describe the regression model, and the hiLasso estimation problem in Section 3.2. In Section 3.3, we discuss techniques for proving consistency of high-dimensional estimators, and prove the statistical consistency of hiLasso. Finally, we briefly describe some efficient optimization methods for solving the hiLasso problem in Section 3.4.

3.2 Problem Statement

3.2.1 Regression Model

Consider a linear statistical model defined as:

$$\mathbf{y} \sim \mathbf{X}\theta^* + w, \tag{3.1}$$

where $y \in \mathbb{R}^n$ is an n-dimensional vector of observations of a response variable, $\theta^* \in \mathbb{R}^p$ is the coefficient vector associated with p covariates, $X \in \mathbb{R}^{n \times p}$ is the covariate or design matrix, and $w \in \mathbb{R}^n$ is the noise vector. Our goal is two-fold:

1. understand which covariates are relevant/important for predicting the response variable, and
2. build a suitable regressor based on these relevant variables.

Assuming that the noise vector w follows a Gaussian distribution, estimating the vector θ^* amounts to solving the "ordinary least squares" (OLS) problem:

$$\hat{\theta}_{OLS} = \underset{\theta \in \mathbb{R}^p}{\arg,\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\theta\|_2^2 \right\}. \tag{3.2}$$

Clearly, when $n < p$, the system is unidentifiable and we will obtain multiple solutions $\hat{\theta}_{OLS}$. Moreover, in general, all coefficients of $\hat{\theta}_{OLS}$ will be non-zero, signifying statistical dependency

of the "response" variable on all covariates. As is well known in statistical literature [26], the OLS estimate has large variance and hence, is not robust. Also, the estimate is not interpretable in terms of the particular application domain under consideration due to the presence of many spurious dependencies.

In such cases, a regularizer $r(\theta)$ is added to the squared loss function in order to have a more robust estimate of θ^* [26]. In many applications, such as climate, the dependencies are, in general, *sparse*, meaning that most of the coefficients of $\hat{\theta}$ are zero [99, 100]. To promote sparsity in the estimate, sparsity-inducing convex norm regularizers are commonly used [27, 101]. These sparse methods offer significant computational benefits over traditional feature selection methods and some have been proven to be statistically consistent [59, 101].

As an example, consider the problem of predicting land climate variables using variables measured over oceans. The covariates in our problem are multiple climate variables measured at different geographical locations. This spatial structure of the data indicates a natural "grouping" of the variables at each ocean location. Simple sparse regularizers, such as the LASSO penalty [27] do not respect this structure inherent in the data. Therein arises the need to have regularizers which impose *structured sparsity* that respects this spatial nature. The model that we use incorporates such a regularizer and is called *Hierarchical Lasso*, or hiLasso[32]. The next subsection describes the model.

3.2.2 The hiLasso

HiLasso solves a **regularized** estimation problem which utilizes a convex norm regularization function that enforces structured sparsity. The structure enforced by this norm is encoded by a hierarchical tree within the dimensions $\{1, \ldots, p\}$. This norm is called a tree norm and is defined as follows. Consider a grouping $\mathcal{G}$ if the dimensions $\{1, \ldots, p\}$, where the groups form a tree-hierarchy, such that groups at each level of the tree are disjoint, and sibling groups at each level share only one parent.

Note that the root node consisting of all the dimensions $\{1, ..., p\}$ is not considered here. Let the height of the tree be $h+1$, with the leaves having a height 0 and the root having a height $h+1$. Let the maximum size of a group at height i be m_i. Let the nodes (groups) at height i be denoted by $\mathcal{G}^i := \{G_j^i\}$, $j = 1, \ldots, n_i$. Note that $n_0 = p$ and $m_0 = 1$. The group norm at

height i is computed as:

$$\|\theta\|_{(\mathcal{G}^i,\nu)} := \sum_{j=1}^{n_i} \|\theta_{G_j^i}\|_\nu \tag{3.3}$$

For any $(\alpha_0, \alpha_1, \ldots, \alpha_h)$ such that $1 > \alpha_i > 0\ , \forall i$ and $\alpha_0 + \alpha_1 + \ldots + \alpha_h = 1$, the tree-norm regularizer is formally defined as

$$r(\theta) := r_{tree}(\theta) := \sum_{i=0}^{h} \alpha_i \|\theta\|_{(\mathcal{G}^i,\nu)}\ . \tag{3.4}$$

The regularized estimation problem that hiLasso solves is of the form

$$\hat{\theta} = \underset{\theta\in\mathbb{R}^p}{\arg,\min} \left\{ \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda r(\theta) \right\}, \tag{3.5}$$

3.2.3 Special Cases of hiLasso

Many popular sparse regression problems are special cases of the hiLasso. The L_1-regularized least squares estimation problem, known as **LASSO** [27], is a special case where $h = 0$, and the groups at the leaves are singletons $\{1, \ldots, p\}$. In this case, the regularizer r is simply the L_1 norm of the coefficient θ and the regularized estimation problem is of the form:

$$\hat{\theta}_{Lasso} = \underset{\theta\in\mathbb{R}^p}{\arg,\min} \left\{ \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|_1 \right\} \tag{3.6}$$

The **Group LASSO** [28] is another special case where $h = 0$, and the leaves of the tree are non-overlapping groups formed from the set $\{1, \ldots, p\}$. In this case, the problem takes the form

$$\hat{\theta}_{gpLasso} = \underset{\theta\in\mathbb{R}^p}{\arg,\min} \left\{ \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|_{1,\mathcal{G}} \right\}\ , \tag{3.7}$$

where

$$\|\theta\|_{1,\mathcal{G}} = \sum_{k=1}^{T} \|\theta_{G_k}\|_2 \tag{3.8}$$

is the group Lasso norm.

For a two-level tree hierarchy, the regularizer r consists of a convex combination of a group lasso norm [28], and an L_1 norm. This estimator is called the **Sparse Group Lasso (SGL)**, and is defined as

$$\hat{\theta}_{SGL} = \underset{\theta\in\mathbb{R}^p}{\arg,\min} \left\{ \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda r(\theta) \right\}, \tag{3.9}$$

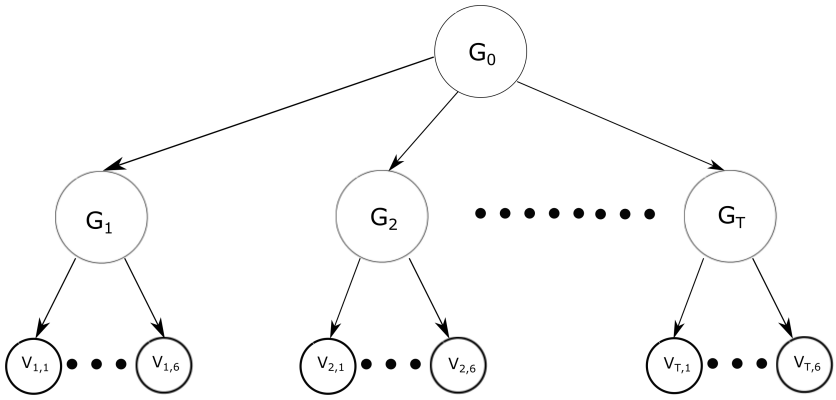


Figure 3.1: The hierarchical structure of the SGL norm. $G_1, \ldots, G_T$ are groups of variables, where $V_{k,1} \ldots, V_{k,m}$ are variables in G_k.

where r is the SGL regularizer given by

$$r(\theta) := r_{(1,\mathcal{G}_2,\alpha)} = \alpha\|\theta\|_1 + (1-\alpha)\|\theta\|_{1,\mathcal{G}} \ , \tag{3.10}$$

where $\mathcal{G} = \{G_1, \ldots, G_T\}$ are the groups of variables (Fig. 3.1. The mixed norm $\|\theta\|_{1,\mathcal{G}}$ penalizes groups of variables, while the L_1 norm $\|\theta\|_1$ promotes sparsity among variables within each group.

In the next section, following the analysis technique developed in [30], we prove that, under fairly general conditions, hiLasso is statistically consistent in estimating the true parameter θ^* of the distribution from which the data samples (X, y) were generated. For the special case of SGL, we provide explicit bounds for the consistency of SGL.

3.3 Consistency of Hierarchical Lasso

3.3.1 Formulation

Let $Z_1^n := \{Z_1, \ldots, Z_n\}$ denote n observations drawn i.i.d. according to some distribution $\mathbb{P}$, and suppose that we are interested in estimating the parameter θ of the distribution $\mathbb{P}$. Let $\mathcal{L}(\theta; Z_1^n) = \|\mathbf{y} - \mathbf{X}\theta\|^2$ be the squared loss function. Given the hiLasso regularizer $r : \mathbb{R}^p \mapsto \mathbb{R}$, the hiLasso estimator is given by

$$\hat{\theta}_{\lambda_n} \in \underset{\theta \in \mathbb{R}^p}{\arg,\min} \ \{\mathcal{L}(\theta; Z_1^n) + \lambda_n r(\theta)\} \ , \tag{3.11}$$

where $\lambda_n > 0$ is a user-defined regularization penalty.

In this work, we assume that the noise vector w is zero mean and has sub-Gaussian tails, i.e., there is a constant $\sigma > 0$ such that for any $v, \|v\|_2 = 1$, we have

$$\mathbb{P}(|\langle v, w\rangle| \geq \delta) \leq 2\exp\left(-\frac{\delta^2}{2\sigma^2}\right) , \quad \text{for all } \delta > 0 . \tag{3.12}$$

The condition holds in the special case of Gaussian noise; it also holds whenever the noise vector w consists of independent bounded random variables.

3.3.2 Assumptions on Regularizer and Loss Function

Following [30], the first key requirement for the analysis is a property of the regularizer r. Let us assume that the optimal statistical parameter θ^* lies in a subspace A of dimensionality s_A. Let $A^\perp$ be the orthogonal subspace of the space A. The regularizer r is defined to be *decomposable* w.r.t. the subspace pair $(A, A^\perp)$ if, for any $\alpha \in A$ and $\beta \in A^\perp$,

$$r(\alpha + \beta) = r(\alpha) + r(\beta) . \tag{3.13}$$

Let us define the error vector $\hat{\Delta}_{\lambda_n} := \hat{\theta}_{\lambda_n} - \theta^*$, and the projection operator $\Pi_A : \mathbf{R}^p \mapsto A$, such that

$$\Pi_A(u) = \arg, \min_{v \in A} \|u - v\|_* ,$$

for some given error norm $\|\cdot\|_*$. [30] showed that for a decomposable regularizer r, and choosing λ_n satisfying

$$\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n)) ,$$

the error $\hat{\Delta}$ lies in the set

$$\mathbb{C}(A; \theta^*) := \{\Delta \in \mathbf{R}^p | r(\Pi_{A^\perp}(\Delta)) \leq 3r(\Pi_A(\Delta))\} , \tag{3.14}$$

The second key requirement, as stated in [30], is that the loss function $\mathcal{L}$ should satisfy the Restricted Strong Convexity (RSC) property. Let us define $\delta\mathcal{L}(\Delta, \theta^*) := \mathcal{L}(\theta^* + \Delta; Z_1^n) - \mathcal{L}(\theta^*; Z_1^n) - \langle\nabla\mathcal{L}(\theta^*; Z_1^n), \Delta\rangle$. $\mathcal{L}$ satisfies RSC with curvature $\kappa_\mathcal{L} > 0$ and tolerance function $\tau_\mathcal{L}$ if, for all $\Delta \in \mathbb{C}(A, B; \theta^*)$,

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_\mathcal{L}\|\Delta\|_*^2 - \tau_\mathcal{L}^2(\theta^*) . \tag{3.15}$$

Further, [30] defines a subspace compatibility constant with respect to the pair $(r, \|\cdot\|_*)$ for any subspace $B \subseteq \mathbb{R}^p$ as follows:

$$\Psi(B) := \sup_{u \in B \setminus \{0\}} \frac{r(u)}{\|u\|_*} . \tag{3.16}$$

For the squared loss, it can be shown that with high probability, $\delta\mathcal{L}$ satisfies

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_1 \|\Delta\|_*^2 - \kappa_2 \, g(n,p) r^2(\Delta), \;\; \forall \, \|\Delta\|_* \leq 1 , \tag{3.17}$$

where $g(n,p)$ is a function of the sample size n and dimensionality p . As shown in [30] , this implies restricted strong convexity under mild conditions on the sample size. We would illustrate in Section 3.3.3 that this form of RSC holds for hiLasso.

Based on the assumption that the norm regularizer r is decomposable w.r.t. a subspace pair $(A, A^{\perp})$, [30] presents the following key result:

Theorem 1 *Consider the convex program in (3.11) based on a strictly positive regularization constant*

$$\lambda_n \geq 2r^*(\nabla\mathcal{L}(\theta^*; Z_1^n)) . \tag{3.18}$$

Then any optimal solution $\hat{\theta}_{\lambda_n}$ to (3.11) satisfies the bound

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_*^2 \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(B) + \frac{2\lambda_n}{\kappa_{\mathcal{L}}}\tau_{\mathcal{L}}^2(\theta^*) . \tag{3.19}$$

3.3.3 Analysis for hiLasso

Our analysis consists of three key parts: (i) Showing that the regularizer r_{tree} is decomposable, (ii) Showing that the loss function satisfies the RSC condition, and (iii) Choosing a λ_n which satisfies the prescribed lower bound.

Following [30], we assume that for each $k = 1, \ldots, p$

$$\frac{\|X_k\|_2}{\sqrt{n}} \leq 1 . \tag{3.20}$$

Note that the assumption can be satisfied by simply rescaling the data, and is hence without loss of generality. Further, the above assumption implies that

$$\frac{\|X_{G_j^i}\|_{\nu \to 2}}{\sqrt{n}} \leq 1 , \tag{3.21}$$

where the operator norm

$$\|X_{G_t}\|_{\nu\to 2} := \max_{\|\theta\|_\nu=1} \|X_{G_t}\theta\|_2 \ .$$

Decomposability of Regularizer

We may note that the group norm at a particular height in the tree, $\|\theta\|_{(\mathcal{G}^i,\nu)}$ is over groups which are disjoint. Hence it decomposes over the subspace spanned by each group. Therefore, following the definitions and arguments in [30], the tree-norm is decomposable.

Restricted Strong Convexity

As shown in [30] , if X is formed by sampling each row $X_i \sim N(0,\Sigma)$, referred to as the Σ-*Gaussian ensemble*, then there exists constants $(\kappa_{1,i},\kappa_{2,i})$, such that, with high probability

$$\frac{\|X\theta\|_2^2}{n} \geq \kappa_{1,i}\|\theta\|_2^2 - \kappa_{2,i}\|\theta\|_{(\mathcal{G}^i,\nu)}^2 \ , \tag{3.22}$$

for groups at height i . Now, from the definition of the hierarchical tree structure norm, we have

$$\begin{aligned} r^2(\theta) &= \left(\sum_i \alpha_i \|\theta\|_{(\mathcal{G}^i,\nu)}\right)^2 \\ &= \sum_i \alpha_i^2 \|\theta\|_{(\mathcal{G}^i,\nu)}^2 + \sum_{i,j:i\neq j} \alpha_i\alpha_j \|\theta\|_{(\mathcal{G}^i,\nu)}\|\theta\|_{(\mathcal{G}^j,\nu)} \\ &\geq \sum_i \alpha_i^2 \|\theta\|_{(\mathcal{G}^i,\nu)}^2 \ , \end{aligned} \tag{3.23}$$

where the inequality follows from the non-negativity of α and group norms.

From (3.22) , we can conclude that with high probability,

$$\begin{aligned} &\left(\sum_i \alpha_i^2\right)\frac{\|X\theta\|_2^2}{n} \\ &\geq \left(\sum_i \alpha_i^2\kappa_{1,i}\right)\|\theta\|_2^2 - (\max_i \kappa_{2,i})\left(\sum_i \alpha_i^2\|\theta\|_{(\mathcal{G}^i,\nu)}^2\right) \\ \Rightarrow \quad &\frac{\|X\theta\|_2^2}{n} \geq \kappa_1\|\theta\|_2^2 - \kappa_2 r^2(\theta) \ , \end{aligned} \tag{3.24}$$

and RSC follows from the discussion in Section 3.3.2 .

Bounds for λ_n

Recall from Theorem 1 that the λ_n needs to satisfy the following lower bound:

$$\lambda_n \geq 2r^*_{tree}(\nabla \mathcal{L}(\theta^*; Z_1^n)) \, . \tag{3.25}$$

A key issue with the above lower bound is that it is a random variable depending on Z_1^n. A second issue is that the conjugate r^*_{tree} for the mixed norm $r_{tree}(v)$ may not be obtainable in closed (non-variational) form. So we first obtain an upper bound $\bar{r}^*_{tree}$ on r^*_{tree}, and choose a λ_n which will satisfy the lower bound in (3.25) with high probability over choices of Z_1^n.

By definition

$$\begin{aligned}
& r^*_{tree}(v) \\
& \quad = \sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u, v \rangle}{r_{tree}(u)} \\
& \quad = \sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u, v \rangle}{\sum_{i=0}^h \alpha_i \|u_{\mathcal{G}^i}\|_{(1,\nu)}} \\
& \quad \overset{(a)}{\leq} \sup_{u \in \mathbb{R}^p \setminus \{0\}} \left[\sum_{i=0}^h \alpha_i \frac{\langle u, v \rangle}{\|u_{\mathcal{G}^i}\|_{(1,\nu)}} \right] \\
& \quad \leq \sum_{i=0}^h \alpha_i \sup_{u \in \mathbb{R}^p \setminus \{0\}} \frac{\langle u, v \rangle}{\|u_{\mathcal{G}^i}\|_{(1,\nu)}} \\
& \quad = \sum_{i=0}^h \alpha_i r^*_{\mathcal{G}^i_\nu}(v) = \bar{r}^*_{tree}(v) \, ,
\end{aligned} \tag{3.26}$$

where (a) follows from Jensen's inequality and $r^*_{\mathcal{G}^i_\nu}$ is the conjugate norm of $r_{\mathcal{G}^i_\nu}(v) = \sum_{j=1}^{n_i} \|v_{G^i_j}\|_\nu$ given by

$$r^*_{\mathcal{G}^i_\nu}(v) = \max_{j=1,\ldots,n_i} \|v_{G^i_j}\|_{\nu^*} \, , \tag{3.27}$$

where $\nu^* > 0$ satisfies $\frac{1}{\nu} + \frac{1}{\nu^*} = 1$.

By definition, we have $\nabla L(\theta^*; Z_1^n) = \frac{X^T w}{n}$ where $w = y - X\theta^*$ is a zero mean sub-Gaussian random variable. As a result, it is sufficient to choose λ_n satisfying:

$$\lambda_n \geq 2 \left[\sum_{i=0}^h \alpha_i \left(\max_{j=1,\ldots,n_i} \left\| \frac{X^T_{G^i_j} w}{n} \right\|_{\nu^*} \right) \right] . \tag{3.28}$$

We make use of the following Lemma in our subsequent analysis for choosing λ_n appropriately.

Lemma 1 *At any height i of the tree, with probability at least $1 - 2\exp(-\frac{n\delta^2}{2\sigma^2})$, we have*

$$\max_{j=1,\ldots,n_i} \left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} \leq 2\sigma \frac{m_i^{1-\mathbb{1}/\nu}}{\sqrt{n}} + \delta \,, \tag{3.29}$$

where m_i is the size of the largest group at height i.

Proof: For any $j \in \{1, \ldots, n_i\}$ and $i \in \{0, \ldots, h\}$, consider the random variable:

$$Y_j^i = Y_j^i(w) = \left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} .$$

Following exactly similar arguments as in [30], we can show that $Y_j^i(w)$ is a Lipschitz function of w with constant $\frac{1}{\sqrt{n}}$. It follows that

$$\mathbb{P}\left[Y_j^i \geq E[Y_j^i] + \delta\right] \leq 2\exp(-\frac{n\delta^2}{2\sigma^2}) \,. \tag{3.30}$$

Suitably applying the Sudakov-Fernique comparison principle [102, 103] shows that:

$$E[Y_j^i] = E\left[\left\| \frac{X_{G_j^i}^T w}{n} \right\|_{\nu^*} \right] \leq 2\sigma \frac{m_i^{1-\mathbb{1}/\nu}}{\sqrt{n}} \,, \tag{3.31}$$

so that we have

$$P\left[Y_j^i \geq 2\sigma \frac{m_i^{1-\mathbb{1}/\nu}}{\sqrt{n}} + \delta \right] \leq 2\exp(-\frac{n\delta^2}{2\sigma^2}) \,, \tag{3.32}$$

for each $j \in \{1, \ldots, n_i\}$. Hence, we have

$$P\left[\max_{j=1,\ldots,n_i} Y_j^i \geq 2\sigma \frac{m_i^{1-\mathbb{1}/\nu}}{\sqrt{n}} + \delta \right] \leq 2\exp(-\frac{n\delta^2}{2\sigma^2}) \,. \tag{3.33}$$

■

Next we use Lemma 1 to obtain a choice of λ_n such that (3.28) holds with high probability. Multiplying both sides of (3.33) by α_i and applying the union bound over $i = 0, \ldots, h$ we obtain

$$\begin{aligned} \mathbb{P}\left\{ \sum_{i=0}^{h} \alpha_i \left(\max_{j=1,\ldots,n_i} Y_j^i \right) \geq \; 2\sigma \frac{\sum_{i=0}^{h} \alpha_i m_i^{1-1/\nu}}{\sqrt{n}} + \delta \right\} \\ \leq 2\exp\left[-\frac{n\delta^2}{2\sigma^2} + \log(h+1) \right] \,. \end{aligned} \tag{3.34}$$

For any $k > 0$, choosing

$$\delta = \sigma \sqrt{\frac{2(k+1)\log(h+1) + k \log p}{n}}\,, \tag{3.35}$$

we get the following result:

Lemma 2 *If*

$$\lambda_n \geq 2\sigma \left\{ \frac{2\sum_{i=0}^{h} \alpha_i m_i^{1-1/\nu}}{\sqrt{n}} + \sqrt{\frac{2(k+1)\log(h+1) + k\log p}{n}} \right\},$$

then

$$\mathbb{P}\left[\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*; Z_1^n))\right] \geq 1 - \frac{2}{p^k (h+1)^k}\,.$$

3.3.4 Analysis of Sparse Group Lasso

The Sparse Group Lasso is a special case of hiLasso, when the height of the tree is 2. The first level of the tree contains nodes corresponding to the T disjoint groups $\mathcal{G} = \{G_1, \ldots, G_T\}$, while the second level contains the singletons. It combines a group-structured norm with a element-wise norm (3.10). For ease of exposition, we assume the groups G_t are of the same size, say of m indices, so we have T groups of size m, and $p = Tm$.

A direct analysis for SGL using the proof method just described provides the following lemma:

Lemma 3 *If*

$$\lambda_n \geq 2\sigma \left\{ \frac{2(1 + m^{1-\mathbb{1}/\nu})}{\sqrt{n}} + \sqrt{\frac{k \log p}{n}} \right\}, \tag{3.36}$$

then

$$\mathbb{P}\left[\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*; Z_1^n))\right] \geq 1 - \frac{2}{p^k}\,. \tag{3.37}$$

3.3.5 Main Result

A direct application of Theorem 1 now gives the following result:

Theorem 2 *Let A be any subspace of $\mathbb{R}^p$ of dimension s_A which contains θ^*, the optimal (unknown) regression parameter. Then, if λ_n satisfies the lower bound in Lemma 2, with probability at least $\left(1 - \frac{2}{p^k(h+1)^k}\right)$, we have*

$$\|\hat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{9\lambda_n^2}{k_{\mathcal{L}}^2} s_A = O\left(\frac{\log p}{n}\right) , \tag{3.38}$$

where $\hat{\theta}_{\lambda_n}$ is the hiLasso estimator.

Note that the bound on the error scales as the logarithm of the problem dimensionality and inversely as the sample. The scale of the bound is consistent with known rates in sparse regression [30]. In the next section, we discuss some optimization methods for sparse group lasso.

3.4 Optimization Method

Our analysis in the previous section illustrates that SGL encodes a tree-structured hierarchy in grouping covariates which leads to sparsity at two levels: groups and singletons. The different sparsity structures induced by hierarchical norms have been explored in [98] and [65]. The authors have independently proposed methods for optimization.

We follow the proximal method proposed by [65] which solves the primal SGL problem. Note that (3.9) is the sum of two convex functions [104], where the squared loss $\mathcal{L}$ is smooth and the regularizer r is non-smooth. At each iteration t, the proximal algorithm computes the following updates

$$\begin{aligned} \theta_{t+\frac{1}{2}} &= \theta_t - \eta_t \nabla \mathcal{L} \\ \theta_{t+1} &= \arg,\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|\theta - \theta_{t+\frac{1}{2}}\|_2^2 + \lambda_n r_{tree}(\theta) , \end{aligned} \tag{3.39}$$

where η_k is the learning rate. [65] shows that the proximal step can be performed efficiently by one pass through all the groups encoded by the tree hierarchy. The algorithm is initialized at the leaves of the tree and terminates at the root. In essence, the proximal step involves successive

projections into the subspaces spanned by the groups $\mathcal{G}^i$ at height i . As mentioned earlier, these groups $\mathcal{G}^i$ are mutually disjoint. Hence, the algorithm simply performs iterative shrinkage at each step. Since SGL constitutes of a depth-2 hierarchical tree, the proposed algorithm is expected to be fast. Theoretically, it achieves a global convergence rate of $O(\frac{1}{k})$ after k iterations [65].

The authors of [65] have done an efficient implementation of their algorithm in a MATLAB interfaced module called SLEP [66].

3.5 Conclusion

In this chapter, we described the hierarchical tree-structured norm regularizer, and estimation with hiLasso. We illustrated that the Sparse Group Lasso is a special case of hiLasso. We proved non-asymptotic rates of consistency for hiLasso, and showed that the scale of derived error bounds is similar to existing bounds for sparse regression. Finally, we discussed some efficient optimization methods for hiLasso using proximal methods.

Chapter 4

Generalized Dantzig Selector

4.1 Introduction

The Dantzig Selector (DS) [35, 34] provides an alternative to regularized regression approaches such as Lasso [27, 59] for sparse estimation. Although DS does not consider a regularized maximum likelihood approach, [35] has established clear similarities between the estimates from DS and Lasso. While norm regularized regression approaches have been generalized to more general norms, such as decomposable norms [30], the literature on DS has primarily focused on the sparse L_1 norm case, with a few notable exceptions which have considered extensions to sparse group-structured norms [67].

In this chapter, we consider linear models of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \mathbf{w}$, where $\mathbf{y} \in \mathbb{R}^n$ is a set of observations, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix, and $\mathbf{w} \in \mathbb{R}^n$ is i.i.d. noise. For *any* given norm $\mathcal{R}(\cdot)$, the parameter $\boldsymbol{\theta}^*$ is assumed to structured in terms of having a low value of $\mathcal{R}(\boldsymbol{\theta}^*)$. For this setting, we propose the following Generalized Dantzig Selector (GDS) for parameter estimation:

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg,\min}\ & \mathcal{R}(\boldsymbol{\theta}) \\ \text{s.t. } & \mathcal{R}^*\big(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\big) \leq \lambda_p \ , \end{aligned} \tag{4.1}$$

where $\mathcal{R}^*(\cdot)$ is the dual norm of $\mathcal{R}(\cdot)$, and λ_p is a suitable constant. If $\mathcal{R}(\cdot)$ is the L_1 norm, (4.1) reduces to standard DS [34]. A key novel aspect of GDS is that the constraint is in terms of the dual norm $\mathcal{R}^*(\cdot)$ of the original structure inducing norm $\mathcal{R}(\cdot)$. It is instructive to contrast GDS with the recently proposed atomic norm based estimation framework [69] which, unlike

GDS, considers constraints based on the L_2 norm of the error $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2$, and focuses only on atomic norms.

In this chapter, we consider statistical aspects of the GDS. We establish non-asymptotic high-probability bounds on the estimation error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$. Interestingly, the bound depends on the Gaussian width of the unit norm ball of $\mathcal{R}(\cdot)$ as well as the Gaussian width of suitable set where the estimation error belongs [69, 105]. As a non-trivial example of the GDS framework, we consider estimation using the recently proposed k-support norm [106, 107]. We provide upper bounds for the Gaussian widths of the unit norm ball and the error set as needed in the GDS framework, yielding the first statistical recovery guarantee for estimation with the k-support norm.

The rest of the chapter is organized as follows: We establish statistical recovery results for GDS for any norm in Section 4.2. In Section 4.3, we present estimation error bounds for the k-support norm. We present experimental results in Section 4.4 and conclude in Section 4.5. All technical analyses and proofs are in the supplement.

4.2 Statistical Recovery

Our goal is to provide error bounds on $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ between the population parameter $\boldsymbol{\theta}^*$ and the minimizer $\hat{\boldsymbol{\theta}}$ of (4.1). Let the error vector be defined as $\hat{\Delta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. For any set $\Omega \subseteq \mathbb{R}^p$, we would measure the size of this set using its Gaussian width [73, 69], which is defined as $\omega(\Omega) = \mathbf{E}_{\mathbf{g}}\left[\sup_{\mathbf{z}\in\Omega}\langle \mathbf{g}, \mathbf{z}\rangle\right]$, where $\mathbf{g}$ is a vector of i.i.d. standard Gaussian entries. We also consider the error cone $\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)$, generated by the set of possible error vectors Δ and containing the error vector $\hat{\Delta}$, defined as

$$\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) := \text{cone}\left\{\Delta \in \mathbb{R}^p \ : \ \mathcal{R}(\boldsymbol{\theta}^* + \Delta) \leq \mathcal{R}(\boldsymbol{\theta}^*)\right\} \ . \tag{4.2}$$

Note that this set contains a restricted set of directions and does not in general span the entire space of $\mathbb{R}^p$. Further, let $\Omega_{\mathcal{R}} := \{\mathbf{u} : \mathcal{R}(\mathbf{u}) \leq 1\}$. With these definitions, we obtain our main result.

Theorem 3 *Suppose the design matrix* $\mathbf{X}$ *consists of i.i.d. Gaussian entries with zero mean variance 1, and we solve the optimization problem* (4.1) *with*

$$\lambda_p \geq c\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] \ . \tag{4.3}$$

Then, with probability at least $(1 - \eta_1 \exp(-\eta_2 n))$*, we have*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{4c\Psi_{\mathcal{R}}\omega(\Omega_{\mathcal{R}})}{\kappa_{\mathcal{L}}\sqrt{n}} \,, \tag{4.4}$$

where $\omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})$ *is the Gaussian width of the intersection of* $\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)$ *and the unit spherical shell* $\mathbb{S}^{p-1}$*,* $\omega(\Omega_{\mathcal{R}})$ *is the Gaussian width of the unit norm ball,* $\kappa_{\mathcal{L}} > 0$ *is the gain given by*

$$\kappa_{\mathcal{L}} = \frac{1}{n}\left(\ell_n - \omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})\right)^2 \,, \tag{4.5}$$

$\Psi_{\mathcal{R}} = \sup_{\Delta \in \mathcal{T}_{\mathcal{R}}} \mathcal{R}(\Delta)/\|\Delta\|_2$ *is a norm compatibility factor,* ℓ_n *is the expected length of a length* n *i.i.d. standard Gaussian vector with* $\frac{n}{\sqrt{n+1}} < \ell_n < \sqrt{n}$*, and* $c > 1, \eta_1, \eta_2 > 0$ *are constants.*

Remark: The choice of λ_p is also intimately connected to the notion of Gaussian width. Note that for $\mathbf{X}$ i.i.d. Gaussian entries, and $\mathbf{w}$ i.i.d. standard Gaussian vector, $\mathbf{X}^T\mathbf{w} = \|\mathbf{w}\|_2 \left(\mathbf{X}^T \frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right) = \|\mathbf{w}\|_2 \mathbf{z}$ where $\mathbf{z}$ is an i.i.d. standard Gaussian vector. Therefore,

$$\lambda_p \geq c\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] = c\mathbf{E}_{\mathbf{w}}[\|\mathbf{w}\|_2] \cdot \mathbf{E}_{\mathbf{X}}\left[\mathcal{R}^*(\mathbf{X}^T \frac{\mathbf{w}}{\|\mathbf{w}\|_2})\right] \tag{4.6}$$

$$= c\mathbf{E}_{\mathbf{w}}[\|\mathbf{w}\|_2]\mathbf{E}_{\mathbf{z}}\left[\sup_{\mathbf{u}:\ \mathcal{R}(\mathbf{u}) \leq 1} \langle \mathbf{u}, \mathbf{z} \rangle\right] \tag{4.7}$$

$$= c\ell_n \omega\left(\Omega_{\mathcal{R}}\right) \,, \tag{4.8}$$

which is a scaled Gaussian width of the unit ball of the norm $\mathcal{R}(\cdot)$.

4.2.1 Examples

L_1-norm Dantzig Selector

When $\mathcal{R}(\cdot)$ is chosen to be L_1 norm, the dual norm is the L_∞ norm, and (4.1) is reduced to the standard DS, given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg,\min} \|\boldsymbol{\theta}\|_1 \qquad \text{s.t. } \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_\infty \leq \lambda \,. \tag{4.9}$$

For statistical recovery, we assume that $\boldsymbol{\theta}^*$ is s-sparse, i.e., contains s non-zero entries, and that $\|\boldsymbol{\theta}^*\|_2 = 1$, so that $\|\boldsymbol{\theta}^*\|_1 \leq s$. It was shown in [69] that the Gaussian width of the set $(\mathcal{T}_{L_1}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})$ is upper bounded as $\omega(\mathcal{T}_{L_1}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})^2 \leq 2s \log\left(\frac{p}{s}\right) + \frac{5}{4}s$. Also note

that $\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] = \mathbf{E}[\|\mathbf{w}\|_2]\mathbf{E}[\|\mathbf{g}\|_\infty] \leq \sqrt{n}\sqrt{\log p}$, where $\mathbf{g}$ is a vector of i.i.d. standard Gaussian entries [34]. Further, [30] has shown that $\Psi_\mathcal{R} = \sqrt{s}$. Therefore, if we solve (4.12) with $\lambda_p = c\sqrt{n}\log p$, then

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq 4c\frac{\sqrt{s\log p}}{\kappa_\mathcal{L}\sqrt{n}} = \mathbf{O}\left(\sqrt{\frac{s\log p}{n}}\right) \tag{4.10}$$

with high probability, which agrees with known results for DS [35, 34].

Group-sparse norm Dantzig Selector

Next, consider $\mathcal{R}$ to be the group-sparse mixed L_1/L_2 norm

$$\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{(2,\mathcal{G})} = \sum_{G\in\mathcal{G}} \|\boldsymbol{\theta}_G\|_2 \,, \tag{4.11}$$

where $\mathcal{G} := \{G \,:\, G \subseteq \{1,2,\ldots,p\},\ G_i \cap G_j = \phi\,,\ \cup G = \{1,2,\ldots,p\}\}$ is a set of disjoint groups of features. The regularized regression method based on this norm is known as the group lasso [28]. In this case, (4.1) is reduced to

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg,\min} \|\boldsymbol{\theta}\|_{(2,\mathcal{G})} \qquad \text{s.t. } \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_{(2,\mathcal{G},\infty)} \leq \lambda \,, \tag{4.12}$$

where $\|\boldsymbol{\theta}\|_{(2,\mathcal{G},\infty)}$ is the dual of the group sparse norm

$$\|\boldsymbol{\theta}\|_{(2,\mathcal{G},\infty)} = \max_{G\in\mathcal{G}} \|\boldsymbol{\theta}_G\|_2 \tag{4.13}$$

Assuming that $\boldsymbol{\theta}^*$ is $s_\mathcal{G}$ group sparse, i.e. supported by $s_\mathcal{G}$ groups, and $\boldsymbol{\theta}^* = 1$, we obtain $\|\boldsymbol{\theta}^*\|_{(2,\mathcal{G})} \leq s$. Further, we can show that the Gaussian width of the set $(\mathcal{T}_{(2,\mathcal{G})}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})$ is upper bounded as

$$\omega(\mathcal{T}_{(2,\mathcal{G})}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})^2 \leq (\sqrt{2\log(|\mathcal{G}| - s_\mathcal{G})} + \sqrt{m_\mathcal{G}})^2 s_\mathcal{G} + s_\mathcal{G} m_\mathcal{G} \,, \tag{4.14}$$

where $m_\mathcal{G}$ is the size of the largest group. Further, we can also show that

$$\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] \leq 2\sqrt{n}(\sqrt{m_\mathcal{G}} + \log|\mathcal{G}|) \,, \tag{4.15}$$

and $\Phi_\mathcal{R} = \sqrt{s_\mathcal{G}}$. Therefore, if we solve (4.12) with $\lambda_p = c\sqrt{n}(\sqrt{m_\mathcal{G}} + \log|\mathcal{G}|)$, then

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq 4c\frac{\sqrt{s_\mathcal{G}}(\sqrt{m_\mathcal{G}} + \log|\mathcal{G}|)}{\kappa_\mathcal{L}\sqrt{n}} \tag{4.16}$$

with high probability.

4.3 Dantzig Selection with k-support norm

We first introduce some notations. Given any $\boldsymbol{\theta} \in \mathbb{R}^p$, let $|\boldsymbol{\theta}|$ denote its absolute-valued counterpart and $\boldsymbol{\theta}^{\downarrow}$ denote the permutation of $\boldsymbol{\theta}$ with its elements arranged in decreasing order. In previous work [106, 107], the k-support norm is defined as

$$\|\boldsymbol{\theta}\|_k^{sp} = \min \left\{ \sum_{I \in \mathcal{G}^{(k)}} \|v_I\|_2 \ : \ \text{supp}(v_I) \subseteq I, \ \sum_{I \in \mathcal{G}^{(k)}} v_I = \boldsymbol{\theta} \right\} , \tag{4.17}$$

where $\mathcal{G}^{(k)}$ denotes the set of subsets of $\{1, \ldots, p\}$ of cardinality at most k. The unit ball of this norm is the set $C_k = \text{conv}\left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{\theta}\|_0 \leq k, \|\boldsymbol{\theta}\|_2 \leq 1\right\}$. The dual norm of the k-support norm is given by

$$\|\boldsymbol{\theta}\|_k^{sp^*} = \max \left\{ \|\boldsymbol{\theta}_G\|_2 \ : \ G \in \mathcal{G}^{(k)} \right\} = \left(\sum_{i=1}^{k} |\boldsymbol{\theta}|_i^{\downarrow 2} \right)^{\frac{1}{2}} . \tag{4.18}$$

The k-support norm was proposed in order to overcome some of the empirical shortcomings of the elastic net [108] and the (group)-sparse regularizers. It was shown in [106] to behave similarly as the elastic net in the sense that the unit norm ball of the k-support norm is within a constant factor of $\sqrt{2}$ of the unit elastic net ball. Although multiple papers have reported good empirical performance of the k-support norm on selecting highly correlated features, wherein L_1 regularization fails, there exists no statistical analysis of the k-support norm. Besides, current computational methods consider square of k-support norm in their formulation, which might fail to work out in certain cases.

In the rest of this section, we focus on GDS with $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_k^{sp}$ given as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg,\min} \|\boldsymbol{\theta}\|_k^{sp} \qquad \text{s.t. } \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_k^{sp^*} \leq \lambda_p \ . \tag{4.19}$$

We prove statistical recovery bounds for k-support norm Dantzig selection, which hold even for a high-dimensional scenario, where $n < p$.

4.3.1 Statistical Recovery Guarantees for k-support norm

The analysis of the generalized Dantzig Selector for k-support norm consists of addressing two key challenges. First, note that Theorem 3 requires an appropriate choice of λ_p. Second, one needs to compute the Gaussian width of the subset of the error set $\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1}$. For the

k-support norm, we can get upper bounds to both of these quantities. We start by defining some notations. Let $\mathcal{G}^* \subseteq \mathcal{G}^{(k)}$ be the set of groups intersecting with the support of $\boldsymbol{\theta}^*$, and let S be the union of groups in $\mathcal{G}^*$, such that $s = |S|$. Then, we have the following bounds which are used for choosing λ_p, and bounding the Gaussian width.

Theorem 4 *For the k-support norm Generalized Dantzig Selection problem* (4.19)*, we obtain For the k-support norm Generalized Dantzig Selection problem* (4.19)*, we obtain*

$$\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] \leq \sqrt{n}\left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right) \tag{4.20}$$

$$\omega(\Omega_{\mathcal{R}}) \leq \left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right) \tag{4.21}$$

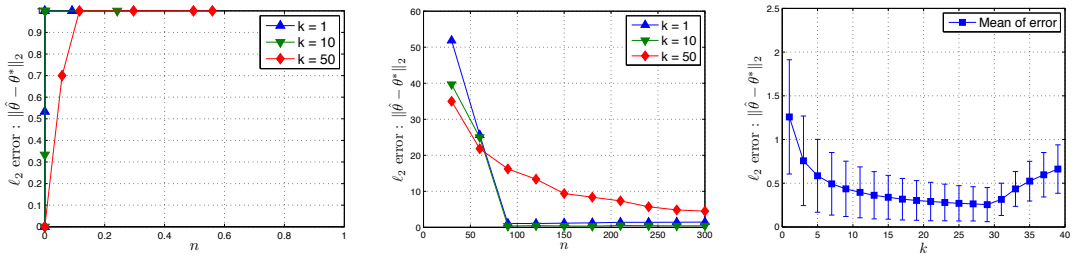
$$\omega(\mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})^2 \leq \left(\sqrt{2k\log\left(p - k - \left\lceil\frac{s}{k}\right\rceil + 2\right)} + \sqrt{k}\right)^2 \cdot \left\lceil\frac{s}{k}\right\rceil + s\,. \tag{4.22}$$

We prove these two bounds using the analysis technique for group lasso with overlaps developed in [105]. Thereafter, choosing $\lambda_p = \sqrt{n}\left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right)$, and under the assumptions of Theorem 3, we obtain the following result on the error bound for the minimizer of (4.19).

Corollary 1 *Suppose that all conditions of Theorem 3 hold, and we solve* (4.19) *with λ_p chosen as above. Then, with high probability, we obtain*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{4c\Psi_{\mathcal{R}}\left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right)}{\kappa_{\mathcal{L}}\sqrt{n}} \tag{4.23}$$

Remark The error bound provides a natural interpretation for the two special cases of the k-support norm, viz. $k = 1$ and $k = p$. First, for $k = 1$ the k-support norm is exactly the same as the L_1 norm, and the error bound obtained will be $\mathrm{O}\left(\sqrt{\frac{s\log p}{n}}\right)$, the same as known results of DS, and shown in Section 2.2. Second, for $k = p$, the k-support norm is equal to the L_2 norm, and the error cone (4.2) is then simply a half space (there is no structural constraint). Therefore, $\Psi_{\mathcal{R}} = O(1)$, and the error bound scales as $\mathrm{O}\left(\sqrt{\frac{p}{n}}\right)$.



(a) ROC curves (b) L_2 error vs. n (c) L_2 error vs. k

Figure 4.1: (a) The true positive rate reaches 1 quite early for $k = 1, 10$. When $k = 50$, the ROC gets worse due to the strong smoothing effect introduced by large k. (b) For each k, the L_2 error is large when the sample is inadequate. As n increases, the error decreases dramatically for $k = 1, 10$ and becomes stable afterwards, while the decrease is not that significant for $k = 50$ and the error remains relatively large. (c) Both mean and standard deviation of L_2 error are decreasing as k increases until it exceeds the number of nonzero entries in $\boldsymbol{\theta}^*$, and then the error goes up for larger k, which matches our analysis quite well. The result also shows that the k-support-norm GDS with suitable k outperforms the L_1 DS when correlated variables present in data (Note that $k = 1$ corresponds to standard DS).

4.4 Numerical Simulations

In this section, we consider the behavior and performance of GDS with k-support norm. For solving (4.19), we use the inexact ADMM algorithm developed in [50]. All experiments are implemented in MATLAB.

Data generation We fixed $p = 600$, and $\boldsymbol{\theta}^* = (\underbrace{10, \ldots, 10}_{10}, \underbrace{10, \ldots, 10}_{10}, \underbrace{10, \ldots, 10}_{10}, \underbrace{0, 0, \ldots, 0}_{570})$ throughout the experiment, in which nonzero entries were divided equally into three groups. The design matrix $\mathbf{X}$ were generated from a normal distribution such that the entries in the same group have the same mean sampled from $\mathcal{N}(0, 1)$. $\mathbf{X}$ was normalized afterwards. The response vector $\mathbf{y}$ was given by $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + 0.01 \times \mathcal{N}(0, 1)$. The number of samples n is specified later.

ROC curves with different k We fixed $n = 400$ to obtain the ROC plot for $k = \{1, 10, 50\}$ as shown in Figure 4.1(a). λ_p ranged from 10^{-2} to 10^3.

L_2 error vs. n We investigated how the L_2 error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ of Dantzig selector changes as the number of samples increases, where $k = \{1, 10, 50\}$ and $n = \{30, 60, 90, \ldots, 300\}$. The plot is shown in Figure 4.1(b).

L_2 error vs. k We also looked at the L_2 error with different k. We again fixed $n = 400$ and varied k from 1 to 39. For each k, we repeated the experiment 100 times, and obtained the mean and standard deviation plot in Figure 4.1(c).

4.5 Conclusion

In this chapter, we introduced the GDS, which generalizes the standard L_1-norm Dantzig Selector to estimation with any norm, such that structural information encoded in the norm can be efficiently exploited. We provide a unified statistical analysis framework for the GDS, which utilizes Gaussian widths of certain restricted sets for proving consistency. In the non-trivial example of k-support norm, our statistical analysis for the k-support norm provides the first result of consistency of this structured norm. Last, experimental results provided sound support to the theoretical development.

Chapter 5

Regression with Noisy Covariates

5.1 Introduction

The study of regression models with error in covariates has gathered some attention in recent literature. In such models we assume that instead of observing $(\mathbf{x}_i, y_i)$ from the linear model $y_i = \langle \boldsymbol{\beta}^*, \mathbf{x}_i \rangle + \epsilon_i$, $(\mathbf{z}_i, y_i)$ is observed, where $\mathbf{z}_i = f(\mathbf{x}_i, \mathbf{w}_i)$ is a noisy version of $\mathbf{x}_i$ corrupted by $\mathbf{w}_i$. Given $\{(\mathbf{z}_i, y_i)\}_{i=1}^n$ we want to compute $\hat{\boldsymbol{\beta}}$, which is l_2 consistent i.e., for the error vector $\Delta = \hat{\beta} - \beta^*$, $\|\Delta\|_2$ is bounded above and the bound shrinks as the number of samples increases.

These models are known as measurement error, errors-in-variables, or noisy covariates and have applications in various areas of science and engineering [109, 110, 111]. The importance of measurement error models is amplified in the era of big data, since large scale and high dimensional data are more prone to noise [111, 112]. In high dimensional setting where $p \gg n$ the classical assumptions required for treatment of measurement error break down [109, 110] and new estimators and methods are required to consistently estimate $\boldsymbol{\beta}^*$. Such challenges have revived measurement error research and several papers have addressed high dimensional issues of those models in recent years [111, 112, 113, 114, 115].

Many recent work reported unstable behavior of standard sparse estimators like LASSO [27] and Dantzig selector (DS) [95] under measurement error which resulted in proposal of new estimators for which some knowledge of noise $\mathbf{w}_i$, and/or $\boldsymbol{\beta}^*$ are required for consistent estimation [111, 112, 113, 114, 115]. Although literature has reached consensus on the inability of classical estimators in dealing with noisy measurements, there is a lack of theoretical results

to describe this phenomenon. In this work, we exactly characterize inconsistency of estimators lacking any knowledge of the measurement noise by showing that the error is bounded by two terms one of which shrinks as the number of samples increases and the other one is irreducible and depends on the covariance of the noise. Our analysis substantially generalizes the existing estimators in the noisy setting, which have only considered sparse regression and l_1 norm regularization.

Most of the work in high dimensional measurement error models assume sparsity on parameter $\boldsymbol{\beta}^*$. However, other structures of $\boldsymbol{\beta}^*$ are of interest in different applications [30, 69]. These structures are formalized as having a small value for $R(\boldsymbol{\beta}^*)$ where R is a suitable norm. Almost none of the previous work in high dimensional measurement error literature has considered structures other than sparsity. In this work, we consider the constrained (DS type) estimators with general norms R, when the design matrix X, with $\mathbf{x}_i$ as its rows, is corrupted by additive independent sub-Gaussian noise matrix W. We study the properties of such estimators where no knowledge of noise W is available. This is in the sharp contrast to the recent literature [111, 112, 114] where the noise covariance $\Sigma_{\mathbf{w}} = \mathbf{E}[W^T W]$ or an estimate of it is required as a part of estimator. [112] uses a maximum likelihood estimator, which always requires estimation of $\Sigma_{\mathbf{w}}$ in order to establish restricted eigenvalue conditions [72, 116, 68] on the estimated sample covariance $\Sigma_{\mathbf{x}}$. [114] showed that for sparse recovery using OMP, although support recovery is possible without any knowledge of $\Sigma_{\mathbf{w}}$, but it is not possible to establish l_2 consistency without estimating $\Sigma_{\mathbf{w}}$ directly. Our analysis shows an explicit characterization of the upper bound on $\|\Delta\|_2$, when $\Sigma_{\mathbf{w}}$ is unknown, which decays as $O(1/\sqrt{n})$ to the variance of the noise. Thus statistical error does not decay to zero, but remains bounded within a norm ball.

Our work advances the understanding of the behavior of high dimensional estimators in the presence of the noise in two key directions. First, our analysis sheds light on the limit of standard estimators under the additive independent sub-Gaussian measurement error model. Second, our results holds for any structured norm R, where DS is a special case when $R(\cdot) = \|\cdot\|_1$. The rest of the chapter is organized as follows. First we introduce the notation and preliminary definitions. Next, we briefly review related work in Section 5.2. In Section 5.3 we formulate the structured estimation problem under noisy designs assumption using constrained optimization and establish non-asymptotic bounds on the error for sub-Gaussian designs and sub-Gaussian noise. Finally, we conclude in Section 5.5.

Table 5.1: Comparison of estimators for design corrupted with additive sub-Gaussian noise

Name	Estimator	Conditions	Bound for $\|\Delta\|_2$
MU [111]	$\min \|\boldsymbol{\beta}\|_1$ s.t. $\|\frac{1}{n}Z^T(\mathbf{y}-Z\boldsymbol{\beta})\|_\infty \le (1+\delta)\delta\|\boldsymbol{\beta}\|_1+\tau$	$\|\frac{1}{n}Z^T\boldsymbol{\epsilon}\|_\infty \le \tau$ $\forall W_{ij}, \|W_{ij}\| \le \delta$	$c\sqrt{s}(\delta+\delta^2)\|\boldsymbol{\beta}^*\|_1 + C\sqrt{\frac{s\log p}{n}}$
IMU [113]	$\min \|\boldsymbol{\beta}\|_1$ s.t. $\|\frac{1}{n}Z^T(\mathbf{y}-Z\boldsymbol{\beta})+\hat{\Sigma}_{\mathbf{w}}\boldsymbol{\beta}\|_\infty \le \mu\|\boldsymbol{\beta}\|_1+\tau$	$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(W_{ij}^2)$ $\Sigma_{\mathbf{w}} = \text{diag}(\sigma_1,\dots,\sigma_p)$ $\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$	$C\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{s\log p}{n}}$
NCL [112]	$\min \frac{1}{2}\boldsymbol{\beta}^T\big(\frac{1}{n}Z^TZ-\hat{\Sigma}_{\mathbf{w}}\big)\boldsymbol{\beta}-\frac{1}{n}\boldsymbol{\beta}^TZ^T\mathbf{y}$ $+\lambda\|\boldsymbol{\beta}\|_1$ s.t. $\|\boldsymbol{\beta}\|_1 \le b_1$	$\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$	$\max\{c\sqrt{s}\lambda, C\|\boldsymbol{\beta}^*\|_2\sqrt{\frac{s\log p}{n}}\}$
NCC [112]	$\min \frac{1}{2}\boldsymbol{\beta}^T\big(\frac{1}{n}Z^TZ-\hat{\Sigma}_{\mathbf{w}}\big)\boldsymbol{\beta}-\frac{1}{n}\boldsymbol{\beta}^TZ^T\mathbf{y}$ s.t. $\|\boldsymbol{\beta}\|_1 \le b_2$	$\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$	$C\|\boldsymbol{\beta}^*\|_2\sqrt{\frac{s\log p}{n}}$
OMP [114]	OMP for recovery of support indecies S: $\hat{\boldsymbol{\beta}}_S = (Z_S^TZ_S-\Sigma_{\mathbf{w}}^S)(Z_S^T\mathbf{y})$	$\mathbf{w}_i \sim \text{Subg}(0,\Sigma_{\mathbf{w}},K_{\mathbf{w}})$ $\forall \beta_i^* \ne 0$ $\|\beta_i^*\| \ge (c\|\boldsymbol{\beta}\|_2+C)\sqrt{\frac{\log p}{n}}$	$(c+C\|\boldsymbol{\beta}^*\|_2)\sqrt{\frac{s\log p}{n}}$

NOTATION AND PRELIMINARIES : We denote matrices by capital letters V, random variables by small letters v and random vectors with bold symbols $\mathbf{v}$. Throughout the chapter c_is and C are positive constants. Consider following norm of random variable v: $\|v\|_{\psi_2} = \sup_{p\ge 1} p^{-1/2}(\mathbb{E}(|v|^p))^{1/p}$. Then v is sub-Gaussian if $\|v\|_{\psi_2} \le K_2$ for a constant K_2. Random vector $\mathbf{v} \in \mathbb{R}^p$ is sub-Gaussian if the one-dimensional marginals $\langle \mathbf{v}, v\rangle$ are sub-Gaussian for all $v \in \mathbb{R}$. The sub-Gaussian norm of $\mathbf{v}$ is defined as $\|\mathbf{v}\|_{\psi_2} = \sup_{v\in S^{p-1}} \|\langle \mathbf{v}, v\rangle\|_{\psi_2}$. We use shorthand $\mathbf{v} \sim \text{Subg}(0,\Sigma_{\mathbf{v}},K_{\mathbf{v}})$ for zero mean sub-Gaussian random vector with covariance $\Sigma_{\mathbf{v}}$ and parameter $K_{\mathbf{v}}$. For any set $A \in \mathbb{R}^p$, the Gaussian width of the set is defined as: $\omega(A) = \mathbb{E}(\sup_{u\in A}\langle g, u\rangle)$, where the expectation is over $g \sim N(0, I_{p\times p})$, a vector of independent zero-mean unit-variance Gaussians.

5.2 Related Work

Over the past decade considerable progress has been made on sparse and structured estimation problems for linear models. Such models assume that the observed pair $(\mathbf{x}_i, y_i)$ follows $y_i = \langle \boldsymbol{\beta}^*, \mathbf{x}_i\rangle + \epsilon_i$, where $\boldsymbol{\beta}^*$ is sparse or suitably structured according to a norm R [69]. In real

world settings, often covariates are noisy, and one observes "noisy" versions $\mathbf{z}_i$ of covariates $\mathbf{x}_i$ corrupted by noise $\mathbf{w}_i$, where $\mathbf{z}_i = f(\mathbf{x}_i, \mathbf{w}_i)$. Two popular model for f are additive, $\mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i$, and multiplicative noise $\mathbf{z}_i = \mathbf{x}_i \circ \mathbf{w}_i$ [111, 112, 114] where $\circ$ is the Hadamard product. Two common choices of $\mathbf{w}_i$ for additive noise case are uniformly bounded [111, 115] and centered subgaussian [112, 114]. In noisy models, a key challenge is to develop estimation methods that are robust to corrupted data, particularly in the high-dimensional regime. Recent work [111, 114] has illustrated experimentally that standard estimators like LASSO and DS perform poorly in the presence of measurement errors. Thus, many recent papers proposed modifications of LASSO, DS or Orthogonal Matching Pursuit (OMP) [111, 112, 114, 113, 115] for handling noisy covariates. However, such estimators may become non-convex [112], or require extra information about optimal $\boldsymbol{\beta}^*$ [112, 114]. Further, most of proposed estimators for sub-Gaussian additive noise require an estimate of the noise covariance $\Sigma_{\mathbf{w}}$ in order to establish statistical consistency [113, 112, 114, 115] or impose more stringent condition, like element-wise boundedness on W [111, 115].

Recent literature on regression with additive measurement error in high dimensions has focused on sparsity, Table 5.1 presents key recent works in this area. The first paper in this line of work [111] introduces matrix uncertainty selector (MU) which belongs to constraint family of estimators. As the first attempt for addressing estimation with measurement error in high dimension, MU imposes restrictive conditions on noise W, namely each element of matrix W needs to be bounded. It worth mentioning that MU does not need any information about noise covariance $\Sigma_{\mathbf{w}}$ and as presented in Table 5.1, it is not consistent, i.e. $c\sqrt{s}(\delta + \delta^2)\|\boldsymbol{\beta}^*\|_1$ term in the upper bound is independent of the number of samples n. This theme repeats in the literature: when $\Sigma_{\mathbf{w}}$ is available proposed estimators are consistent otherwise there is no l_2 recovery guarantee.

Same authors has proposed improved matrix uncertainty selector (IMU) [113] which assumes availability of the diagonal matrix $\hat{\Sigma}_{\mathbf{w}}$ as the covariance of the noise and use it to compensate the effect of the noise. The compensation idea also recurs in the literature where one mitigates $Z^T Z$ by subtracting $\Sigma_{\mathbf{w}}$ and as the result the estimator becomes consistent. Note that both MU and IMU are modification of DS where $\|\boldsymbol{\beta}\|_1$ appears in both constraint and objective of the program. For IMU each row of noise matrix $\mathbf{w}_i$ is sub-Gaussian and independent of $\mathbf{w}_j$, $\mathbf{x}_i$ and ϵ_i and off diagonal of $\Sigma_{\mathbf{w}}$ are zero i.e., W_{ij} are uncorrelated. Following IMU all subsequent work assume sub-Gaussian independent noise and MU and [115] are only estimators that

allows general dependence in noise.

Loh and Wainwright [112] proposed a non-convex modification of LASSO (NCL) along with constraint version of it (NCC) which are equivalent by Lagrangian duality (Table 5.1). In both estimators they substitute the quadratic term $X^T X$ of the LASSO objective with $Z^T Z - \Sigma_{\mathbf{w}}$ which makes the problem non-convex. An interesting aspect of this method is that although a projected gradient algorithm can only reach a local minima, yet any such local minima is guaranteed to have consistency guarantee. Note that for the feasibility of both objectives, [112] requires extra information about the unknown parameter $\boldsymbol{\beta}^*$, particularly b_1 and b_2 should be set to a value greater than $\|\boldsymbol{\beta}^*\|_1$.

In [114], Chen and Caramanis use the OMP [116] for support recovery of a sparse regression problem without knowing the noise covariance. They established non-asymptotic guarantees for support recovery while imposing element-wise lower bound on the absolute value of the support. However, for achieving l_2 consistently, [114] still requires an estimate of the noise covariance $\Sigma_{\mathbf{w}}$, which is in accordance with the requirements of other estimators mentioned above.

Although literature on regression with noisy covariates has only focused on sparsity, the machine learning community recently has made tremendous progress on structured regression that has led to several key publications. [30] provided a general framework for analyzing regularized estimators with decomposable norm of the form $\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; \mathbf{y}, X) + \lambda R(\boldsymbol{\beta})$, and established theoretical guarantees for Gaussian covariates. A number of recent papers [117, 118] have generalized this framework for analyzing estimators with hierarchical structures [33], atomic norms [118] and graphical model structure learning [117]. Recently, [31] established a framework for analyzing regularized estimators with any norm R and sub-Gaussian covariates. On the other hand for constraint estimators [50] has recently generalized the DS for any norm R.

5.3 Noisy Dantzig Selector

We consider the linear model, where covariates are corrupted by additive noise $y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^* \rangle + \epsilon_i, \quad \mathbf{z}_i = \mathbf{x}_i + \mathbf{w}_i$, where $\mathbf{x}_i \sim \text{Subg}(0, \Sigma_{\mathbf{x}}, K_{\mathbf{x}})$, $\epsilon_i \sim \text{Subg}(0, \sigma_\epsilon, K_\epsilon)$ are i.i.d and also independent from one another. Error vector $\mathbf{w}_i \sim \text{Subg}(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$ is independent from both $\mathbf{x}_i$ and ϵ_i. Since $\mathbf{z}_i$ and $\mathbf{x}_i$ are independent, we have $\Sigma_{\mathbf{z}} = \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}}$ and $\mathbf{z}_i \sim \text{Subg}(0, \Sigma_{\mathbf{z}}, K_{\mathbf{z}})$ for $K_{\mathbf{z}} \leq c_1 K_{\mathbf{x}} + c_2 K_{\mathbf{w}}$. In matrix notation, given samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we obtain $\mathbf{y} =$

$X\boldsymbol{\beta}^* + \epsilon$ and $Z = X + W$. The Generalized Dantzig Selector (GDS) with noisy covariates is the following modification of the GDS defined in Chapter 4:

$$\hat{\boldsymbol{\beta}}_c = \arg,\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) \quad \text{s.t.} \quad R^*\left(\frac{1}{n}Z^T\left(\mathbf{y} - Z\boldsymbol{\beta}\right)\right) \leq \lambda_c, \tag{5.1}$$

where $\lambda_c > 0$ is a penalty parameter. R encodes the structure of $\boldsymbol{\beta}^*$. For example, if $\boldsymbol{\beta}^*$ is sparse, i.e. has many zeros, $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and GDS (5.1) is the original Dantzig Selector [34]. When $Z = X$, statistical consistency of GDS has been shown for general norms [50]. However, in the next section, we illustrate that the analysis of [31, 50] can be conducted on GDS with noisy design $Z = X + W$, with similar assumptions, but the resulting estimate is inconsistent.

5.4 Statistical Properties

For noiseless designs, considerable progress has been made in recent years in the analysis of non-asymptotic estimation error $\|\Delta\|_2 = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. Here, we follow the established analysis techniques, while discussing some of the subtle differences in the results obtained due to presence of the noise in covariates. First we discuss the set of directions which contain the error $\boldsymbol{\Delta}$.

Lemma 4 (The Error Set [69, 31]) *For large enough λ_c and for any feasible point $\hat{\boldsymbol{\beta}}_c$ of GDS (5.1) the error vector Δ_c belongs to a restricted error set E_c:*

$$\lambda_c \geq \alpha R^*(\frac{1}{n}Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*)) \Rightarrow E_c = \{\boldsymbol{\Delta}_c : \; R(\boldsymbol{\Delta}_c + \boldsymbol{\beta}^*) \leq R(\boldsymbol{\beta}^*)\}\,. \tag{5.2}$$

We name the cone of E_c as $C_c = Cone(E_c)$.

Proof of the statement is straightforward and only depends on the optimality of $\hat{\boldsymbol{\beta}}$. In the following, we drop the subscript on λ_c, and simply write λ for ease of notation. Next, we discuss the Restricted Eigenvalue (RE) condition on the design matrix that almost all of the high-dimensional consistency analysis relies on [111, 113, 112, 31].

Definition 1 (Restricted Eigenvalue) The design matrix $Z_{n\times p}$ satisfies the restricted eigenvalue condition on the spherical cap $A \subset S^{p-1}$ if $\frac{1}{\sqrt{n}}\inf_{\mathbf{v}\in A}\|X\mathbf{v}\|_2 \geq \kappa > 0$ or in other words, for $\gamma = \sqrt{n}\kappa$:

$$\inf_{\mathbf{v}\in A}\|Z\mathbf{v}\|_2 \geq \gamma > 0\,. \tag{5.3}$$

Intuitively RE condition means that although for $p \gg n$ the matrix Z is not positive definite and the corresponding quadratic form is not strongly convex but in the certain desirable directions represented by A, Z is strongly convex. In GDS the error vector $\mathbf{\Delta}$ directions is given by $A_c = C_c \cap S^{p-1}$.

For noiseless case $Z = X$ when $\mathbf{x}_i$ are Gaussian or sub-Gaussian RE condition is satisfied with high probability after a certain sample size $n > n_0$ is reached, where n_0 determines the sample complexity [31]. Interestingly, recent work has shown that the sample complexity is the square of the Gaussian width of A, $n_0 = O(\omega^2(A))$ [31, 71].

Theorem 5 (Deterministic Error Bound) *Assuming λ satisfies the bound in (5.2) and with sample size $n > n_0$ such that RE condition (5.3) holds over the error direction A, the following deterministic bounds holds for RME and GDS:*

$$\|\mathbf{\Delta}_c\|_2 \leq 4\alpha\Psi(C_c)\frac{\lambda_c}{\kappa}\,, \tag{5.4}$$

where $\Psi(C) = \sup_{\mathbf{u}\in C}\frac{R(\mathbf{u})}{\|\mathbf{u}\|_2}$ is the norm compatibility constant.

Next, we analyze the additive noise case, by obtaining suitable bounds for λ, which sets the scaling of the error bound. Without loss of generality, we will assume $\|\beta^*\|_2 = 1$ for the analysis, noting that the general case follows by a direct scaling of the analysis presented.

5.4.1 Restricted Eigenvalue Condition

For linear models with the square loss function, RE condition is satisfied if (5.3) holds, where $A \subseteq \mathbb{S}^{p-1}$ is a restricted set of directions. Recent literature [31] has proved that the RE condition holds for both Gaussian and sub-Gaussian design matrices. The following theorem shows that RE condition holds for additive noise in measurement with high probability:

Theorem 6 *For the design matrix of the additive noise in measurement $Z = X + W$ where independent rows of X and W are drawn from $\mathbf{x}_i \sim Subg(0, \Sigma_{\mathbf{x}}, K_{\mathbf{x}})$, and $\mathbf{w}_i \sim Subg(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$, and any $\tau > 0$ with probability $1 - 2\exp(-\eta_1\tau^2)$ we have:*

$$\inf_{\mathbf{v}\in A}\|Z\mathbf{v}\|_2 \geq \sqrt{\nu}\sqrt{n} - \eta_0\Lambda_{\max}(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}})\omega(A) - \tau \tag{5.5}$$

where $A \subseteq S^{p-1}$, $\sqrt{\nu} = \inf_{\mathbf{u}\in A}\|(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{w}})^{1/2}\mathbf{u}\|_2$, $\Lambda_{\max}$ is the largest eigenvalue function, and $\eta_0, \eta_1 > 0$ are constants depending on $K_{\mathbf{x}}$ and $K_{\mathbf{w}}$.

5.4.2 Regularization Parameter

The statistical analysis requires $\lambda \geq \alpha R^*(\frac{1}{n}Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*))$. For the noiseless case, we note that $\mathbf{y} - Z\boldsymbol{\beta}^* = \mathbf{y} - X\boldsymbol{\beta}^* = \boldsymbol{\epsilon}$, the noise vector, so that $R^*(\frac{1}{n}Z^T(y - Z\beta^*)) = R^*(\frac{1}{n}X^T\boldsymbol{\epsilon})$. Using the fact that X and $\boldsymbol{\epsilon}$ are sub-Gaussian and independent, recent work has shown that $E[R^*(\frac{1}{n}X^T\epsilon)] \leq \frac{c}{\sqrt{n}}\omega(\Omega_R)$, where $\Omega_R = \{\mathbf{u} \in \mathbb{R}^p | R(\mathbf{u}) \leq 1\}$. For l_1 norm, i.e., LASSO, Ω_R is the unit l_1 ball, and $\omega(\Omega_R) \leq c_2\sqrt{\log p}$. Here we have the following bound on λ:

Theorem 7 *Assume that X and W are zero mean sub-Gaussian matrices. Then,*

$$\mathbf{E}\left[R^*\left(\frac{1}{n}Z^T(\mathbf{y} - Z\boldsymbol{\beta}^*)\right)\right] \leq \nu R(\boldsymbol{\beta}^*) + \frac{C\omega(\Omega_R)}{\sqrt{n}}\,, \tag{5.6}$$

where $\nu = \sup_{\mathbf{u}\in\Omega_R} \|\Sigma_w^{1/2}\mathbf{u}\|_2^2$, and $C > 0$ is a constant dependent on the sub-Gaussian norms of the X and W.

Remark 1: Theorem 7 illustrates that λ does not decay to 0 with increasing sample size, but approaches the operator norm of the covariance matrix Σ_w. Particularly, when the noise W is i.i.d. with variance σ_w^2, the error is bounded above by σ_w^2. **Remark 2:** The main consequence of Theorem 7 is an inconsistency result for the statistical error $\|\Delta\|_2$. We note that in (5.4), when $n > n_0$, κ is a positive quantity that approaches the minimum eigenvalue of $\Sigma_x + \Sigma_w$ with increasing sample size. Therefore, the scaling of λ determines the error bounds. Theorem 7 proves that the error bound can be as small as the variance of the noise. When $W = 0$, consistency rates are exactly the same as the noiseless case.

5.4.3 Consistency with Noise Covariance Estimates

Theorem 7 shows that with no informations about the noise, current analyses can not guarantee statistical consistency for noisy covariates model. At the same time, appearance of $\Sigma_{\mathbf{w}}$ in the upper bound of (5.6), suggests to use noise covariance estimate to reduce the Motivated by this observation and recent line of work, we focused on scenarios in which an estimate of the noise covariance matrix $\hat{\Sigma}_{\mathbf{w}}$ is available, e.g. from repeated measurements Z for the same design matrix X, or from independent samples of W. We follow [112] and assume that independent observation from W is possible, and form $\hat{\Sigma}_{\mathbf{w}} = \frac{1}{n}W_0^T W_0$. Each element of $\hat{\Sigma}_{\mathbf{w}} - \Sigma_{\mathbf{w}}$ is a centered sub-exponential random variable for which $\hat{\Sigma}_w$ concentrates as

$P\left[\|\Sigma_{\mathbf{w}} - \hat{\Sigma}_{\mathbf{w}}\|_{\max} \geq C\sqrt{\frac{\log p}{n}}\right] \leq p^{-c}$ [72, 22]. Having $\hat{\Sigma}_{\mathbf{w}}$ in hand we modify GDS estimator in the following way. Consider the matrix $\hat{\Gamma} = \frac{1}{n}Z^T Z - \hat{\Sigma}_{\mathbf{w}}$ where $\hat{\Sigma}_{\mathbf{w}}$ compensates the effect of noise W, then:

$$\text{Noisy GDS:} \quad \hat{\boldsymbol{\beta}}_c = \arg,\min_{\boldsymbol{\beta}} R(\boldsymbol{\beta}) \quad \text{s.t.} \quad R^*\left(\tfrac{1}{n}Z^T\mathbf{y} - \hat{\Gamma}\boldsymbol{\beta}\right) \leq \lambda_c. \tag{5.7}$$

The following theorems show that the noisy GDS (5.7), rectified with the sample noise covariance $\hat{\Sigma}_{\mathbf{w}}$, provides consistent estimate of $\boldsymbol{\beta}^*$.

Theorem 8 *For the design matrix of the additive noise in measurement $Z = X + W$ where independent rows of X and W are drawn from $\mathbf{x}_i \sim$ Subg$(0, \Sigma_{\mathbf{x}}, K_{\mathbf{x}})$, and $\mathbf{w}_i \sim$ Subg$(0, \Sigma_{\mathbf{w}}, K_{\mathbf{w}})$, and for the noise covariance estimate $\hat{\Sigma}_{\mathbf{w}}$ discussed above, we have the following high probability error bound for the constrained estimator (5.7):*

$$\|\boldsymbol{\Delta}_c\|_2 \leq \frac{4\alpha\Psi(C_c)}{\kappa}\left[\frac{c\omega(\Omega_R)}{\sqrt{n}} + C\sqrt{\frac{\log p}{n}}\right] \tag{5.8}$$

Remark: Note that when R is the vector l_1-norm $\omega(\Omega_R) \leq \sqrt{s\log p}$, and we get the rate of $O(\sqrt{\frac{s\log p}{n}})$ for (5.8) which matches the IMU bound of [113]. The decaying term of the bound in (5.8) comes from λ, the following lemma clarifies this fact:

Lemma 5 (Bound on λ knowing $\hat{\Sigma}_{\mathbf{w}}$) *With high probability, the lower bound $R^*\left(\frac{1}{n}Z^T\mathbf{y} - \hat{\Gamma}\boldsymbol{\beta}^*\right)$ of λ in (5.7) is $\frac{c\omega(\Omega_R)}{\sqrt{n}} + C\sqrt{\frac{\log p}{n}}$.*

5.5 Conclusion

In this work we investigate consistency of the constrained estimators for structured estimation in high dimensional scaling when covariates are corrupted by additive sub-Gaussian noise. Our analysis holds for any norm R, and shows that when no knowledge of noise statistics exists, established methods are inconsistent, but the statistical error can be bounded by the covariance of the noise. Further, when an estimate of the noise covariance is available, the estimator achieves consistent statistical recovery.

Part II

Applications

Chapter 6

Land Variable Regression

6.1 Introduction

Statistical models are being increasingly used to study relationships among climate variables, and develop predictive models based on such relationships. However, climate science problems have some singular challenges, which makes the issue of scientifically meaningful prediction a complex process. Several climate variables are observed at various location on the planet on multiple occasions, thus creating a very large dataset. These variables are dependent between themselves, and across space. However, scientific interpretability and parsimony demands that any discovered relationship among climate variables be simultaneously eclectic and selective. It is not viable to work out such complex dependencies from the first principles of physics, and data mining discovery of potential climate variable relations can be of immense benefit to the climate science community.

The *Sparse Group Lasso* (SGL), discussed in Chapter 3, method is of considerable importance in this context. For a target climate variable in a given location, it allows the selection of other locations that may have an influence through one or more variables, and then allows for a choice of variables at that location. Inherent in this technique is the notion of sparsity, by which only important variables at important locations are selected, from the plethora of potential covariates at various spatial locations.

In this chapter, we discuss the application of sparse modeling to a particular climate prediction task - prediction of land climate variables from measurements of ocean climate variables.

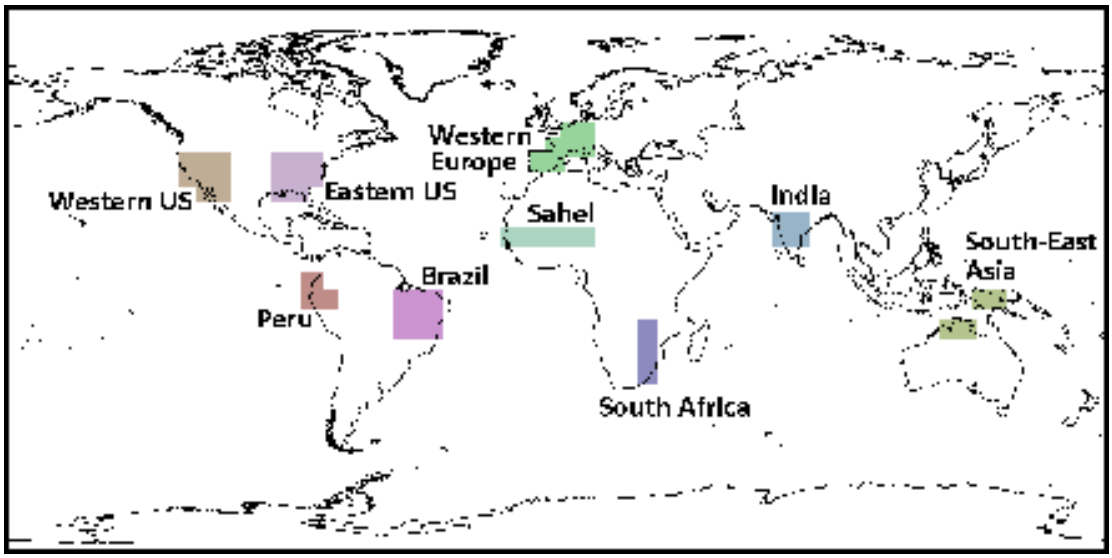


Figure 6.1: Land regions chosen for predictions (picture from [1]).

Through extensive experiments, we illustrate the value of using structured regression for prediction tasks. These experiments consider prediction of mean monthly temperature and mean monthly precipitation over 9 land regions around the world, using covariates which are atmospheric variables measured over the oceans. We show that SGL provides better predictive accuracy and a more interpretable prediction model than the state-of-the-art in climate science. Further, we show that SGL is robust in covariate selection through an empirical analysis of its *regularization path*, and provide climatological insights into the dependencies discovered by SGL.

6.2 Dataset

We begin by describing he dataset used and the preprocessing required for the task. We used the NCEP/NCAR Reanalysis 1 dataset, where we considered the monthly means for 1948-present [51]. The data is arranged as points(locations) on the globe and is available at a $2.5^\circ \times 2.5^\circ$ resolution level. Our main goal is to highlight the utility of using sparse methods to model complex dependencies in climate. Since we trying to model dependencies between target variables and ocean regions, we coarsened the data to $10^\circ \times 10^\circ$ resolution. In total we have data over $N = 756$ time steps. The 6 variables over oceans, considered as covariates, are

(i)Temperature, (ii) Sea Level Pressure, (iii) Precipitation, (iv) Relative Humidity, (v) Horizontal Wind Speed and (vi)Vertical Wind Speed.

We considered 9 "target regions" on land, viz., Brazil, Peru, Western USA, Eastern USA, Western Europe, Sahel, South Africa, Central India and Southeast Asia, as shown in Fig. 6.1. Prediction was done for surface air temperature (SAT) (in $^{\circ}C$) and (b) precipitable water (in kg/m^2) at each of these 9 locations. So, in total, we had 18 response variables. These regions were chosen following [1] because of their diverse geological properties and their impact on human interests.

In total, the dataset contains $L = 439$ locations on the oceans, so that we had $p = 6 \times L = 2634$ covariates in our regression model. We considered the data from January,1948 - December,1997 as the training data and from January,1998 - December,2007 as the test data in our experiments. So, our training set had $n_{train} = 600$ samples and the test set had $n_{test} = 120$ samples. We used the SLEP package [66] for MATLAB to run Sparse Group Lasso on our dataset. It may be noted that we do not take into account temporal relationships that exist in climate data. Moreover, since we consider monthly means, temporal lags of less than a month are typically not present in the data. However, the data does allow us to capture more long-term dependencies present in climate.

6.3 Removing Seasonality and Trend:

As illustrated in [1], seasonality and autocorrelation within climate data at different time points often dominate the signal present in it. Hence, when trying to utilize such data to capture dependency, we look at series of *anomaly* values, i.e., the deviation at a location from the 'normal' value. Firstly, we remove the seasonal component present in the data by subtracting the monthly mean from each data-point and then normalize by dividing it by the monthly standard deviation. At each location we calculate the monthly mean μ_m and standard deviation σ_m for each month $m = 1, \ldots, 12$ (i.e. separately for January, February,... etc.) for the entire time series. Finally, we obtain the anomaly series for location A as the z-score of the variable at location A for month m over the time series.

Further, we need to detrend the data to remove any trend components in the time-series, which might also dominate the signal present in it and bias our regression estimate. Therefore,

Table 6.1: Optimal Choices of (λ_1, λ_2) obtained through 20-fold Cross-Validation.

Region	Variable	λ_1	λ_2
Brazil	Temperature	1	1
	Precipitation	1	1
Peru	Temperature	1	1
	Precipitation	1	1
Western USA	Temperature	1	1
	Precipitation	1	1
Eastern USA	Temperature	1	1
	Precipitation	10	10
Western Europe	Temperature	1	1
	Precipitation	1	1
Sahel	Temperature	1	1
	Precipitation	10	10
South Africa	Temperature	1	1
	Precipitation	10	10
Central India	Temperature	1	1
	Precipitation	1	1
SE Asia	Temperature	1	1
	Precipitation	1	1

we fit a linear trend to the anomaly series at each location over the entire time period 1948-2010 and take the residuals by subtracting the trend. We use this deseasonalized and detrended residuals as the dataset for all our subsequent experiments.

6.4 Choice of penalty parameter (λ):

Table 6.1 shows the optimal choices obtained from cross-validation. The values for different target variables are similar, with the exception of three, which correspond to precipitation in Africa and East USA.

Table 6.2: RMSE scores for prediction of SAT (in $^{\circ}C$) and precipitable water (in kg/m^2) using SGL, LASSO, network clusters [1] and OLS. The number in brackets indicate number of covariates selected by SGL from among the 2634 covariates. The lowest RMSE value in each task is denoted as bold.

Variable	Region	SGL	LASSO	Network Clusters	OLS
Air Temperature	Brazil	**0.198** (651)	0.211	0.534	0.348
	Peru	**0.247** (589)	0.259	0.468	0.387
	West USA	**0.270** (630)	0.291	0.767	0.402
	East USA	**0.304** (752)	0.307	0.815	0.348
	W Europe	0.379 (835)	**0.367**	0.936	0.493
	Sahel	**0.320** (829)	0.322	0.685	0.413
	S Africa	0.136 (685)	**0.130**	0.726	0.267
	India	**0.205** (664)	0.206	0.649	0.3
	SE Asia	0.298 (596)	**0.277**	0.541	0.383
Precipitable Water	Brazil	**0.261** (762)	0.307	0.509	0.413
	Peru	**0.312** (739)	0.344	0.864	0.523
	West USA	**0.451** (824)	0.481	0.605	0.549
	East USA	**0.365** (133)	0.367	0.686	0.413
	W Europe	0.358 (820)	**0.321**	0.450	0.551
	Sahel	0.427 (94)	**0.413**	0.533	0.523
	S Africa	0.235 (34)	**0.215**	0.697	0.378
	India	0.146 (593)	**0.143**	0.672	0.264
	SE Asia	**0.159** (571)	0.168	0.665	0.312

6.5 Prediction Accuracy

Evaluation of our predictions was done by computing the root mean square errors (RMSE) on the test data and comparing the results against those obtained in [1], using OLS estimates, and using Lasso. Note that the problem is high dimensional, since the number of samples (~ 600) for training is much less than the problem dimensionality (~ 2400). [1] uses a correlation based approach to separately cluster each ocean variable into *regions* using a k-means clustering algorithm. The *regions* (78 clusters in total) for all ocean variables are used as covariates for doing linear regression on response variables. Their model is referred to as the *Network Clusters* model. RMSE values were computed, as mentioned earlier, by predicting monthly

mean anomaly for each response variable over the test set for 10 years. The RMSE scores are summarized in Table 6.2. We observed that SGL consistently performs better than both the *Network Clusters* method and the OLS method. The higher prediction accuracy might be explained through the model parsimony that SGL provides. Applying SGL, only the most relevant predictor variables are given non-zero coefficients and any irrelevant variable is considered as noise and suppressed. Since such parsimony will be absent in OLS, the noise contribution is large and therefore the predictions are more erroneous. Further, SGL often performs better than Lasso, although Lasso also provides model parsimony, and has more freedom in selecting relevant covariates. Moreover, structured variable selection provided by SGL often has greater interpretability than Lasso.

The high prediction accuracy of SGL brings to light the inherent power of the model to select appropriate variables (or features) from the covariates during its training phase. To quantitatively elaborate on this aspect, we select two scenarios: (i) Temperature prediction in Brazil and (ii) Temperature prediction in India.

In order to evaluate the covariates which consistently get selected from the set, we repeat the hold out cross-validation experiment with the optimal choices of (λ_1, λ_2) determined earlier for each scenario. During the training phase, an ocean variable was considered *selected*, if it had a corresponding non-zero coefficient. So, in each run of cross-validation, some of the covariates were selected, while others were not. We illustrate our findings in the following subsections.

6.5.1 Region: Brazil

In Fig.6.2, we plot, in descending order of magnitude, the number of times each covariate was selected during cross-validation for temperature prediction in Brazil. We observe that there are ~ 60 covariates among the 2634 covariates that are selected in every single run of cross-validation. In Fig. 6.5, we plot the covariates which are given high coefficient magnitudes by SGL by training on the training dataset from years 1948-1997, in order to illustrate that SGL consistently selects *relevant* covariates. It turns out that these covariates are exactly those which were selected in every cross-validation run. Most of these covariates lie off the coast of Brazil. The influences of horizontal wind speed and pressure is captured, which is consistent with the fact that the ocean currents affect land climate typically through horizontal wind. The tropical climate over Brazil is expected to be influenced by the Inter-tropical Convergence from the north, Polar Fronts from the south, and disturbances in ocean currents from the west, as well as

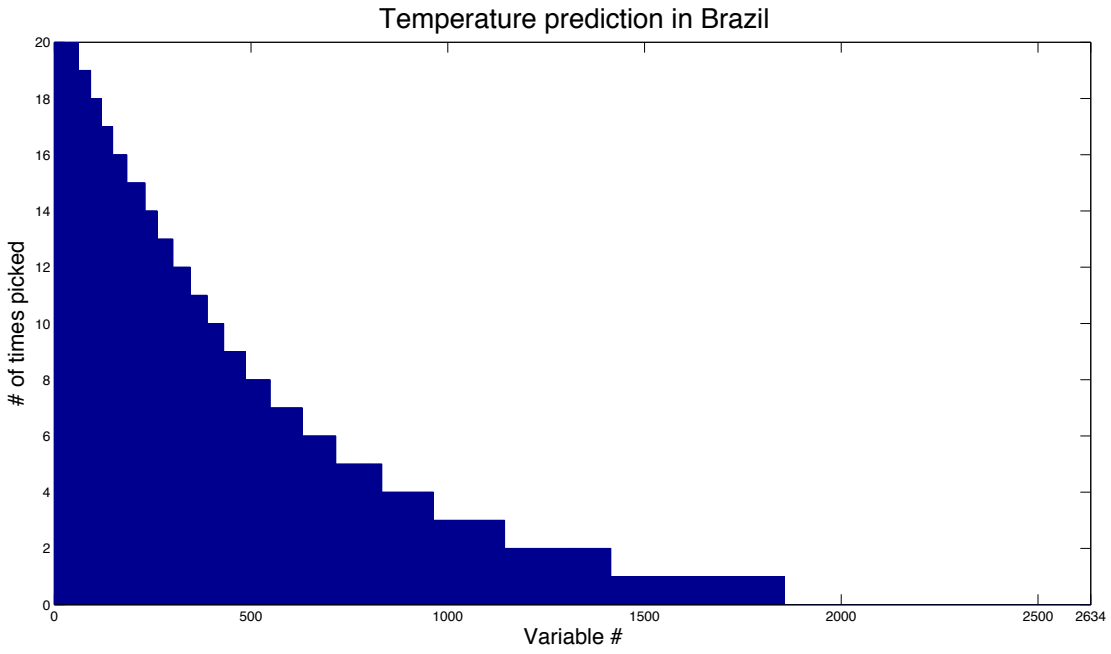


Figure 6.2: Temperature prediction in Brazil: Variables vs. No. of times selected.

the influence of Easterlies from the east and immediate south. It is interesting to see that SGL model captures these influences, as well as the spatial autocorrelation present in climate data, without having any explicit assumptions.

In order to do a comparison, in Fig. 6.6 we plot the variables selected by Lasso in every cross-validation run. There is overlap between this set of variables and the ones selected by SGL, particularly similar variables are selected off the coast of Brazil. However, Lasso has less discretion in the geographic spread of variables chosen. Thus wind, pressure and relative humidity at various locations around the globe are also selected, as shown in the figure. These variables are hard to interpret climatologically, and the model learnt by Lasso, as shown in Table 6.2, often performs worse than SGL

6.5.2 Region: India

Similarly as before, for temperature prediction in India, we construct a histogram of the number of times covariates get selected during cross-validation (Fig. 6.3).

Among the 2634 covariates considered, in this case $\sim$ 65 covariates were chosen in every

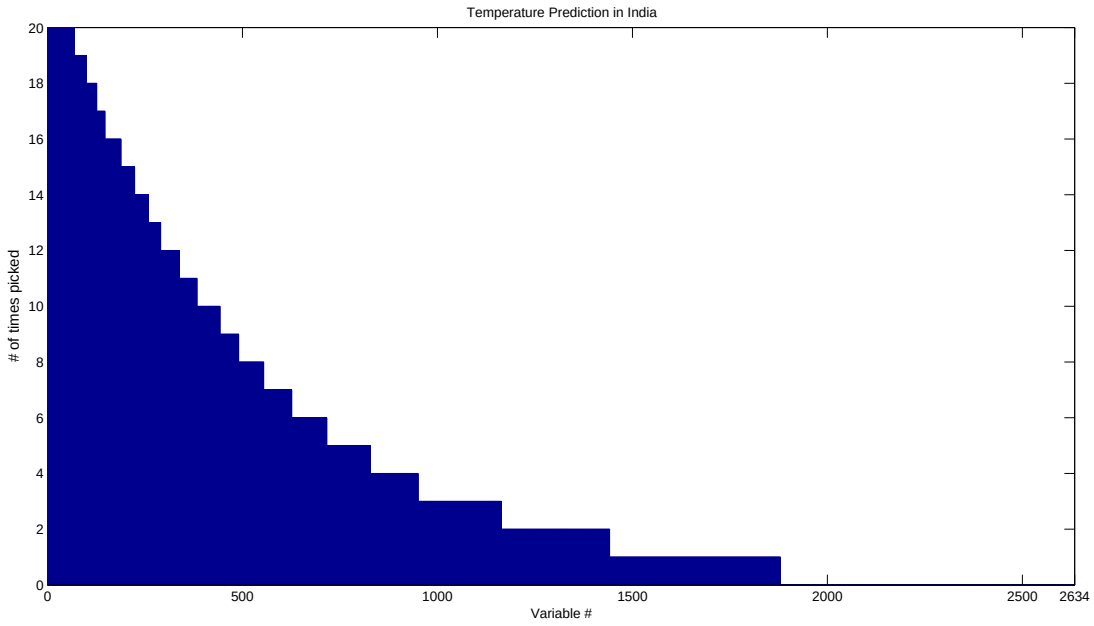


Figure 6.3: Temperature prediction in India: Variables vs. No. of times selected.

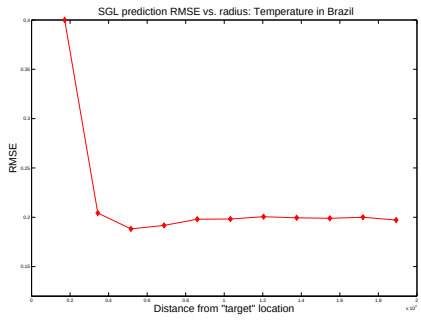
single run of cross-validation. These are plotted in Fig. 6.7. Again, these were the covariates with largest coefficient magnitudes during training on the entire training-set.

We observe the impact of Arabian Sea and Bay of Bengal on the Indian climate. Interestingly, there are some teleconnections which are captured by SGL over the Pacific Ocean, which may be due to the connections between Indian Monsoon and El-Nino [119] and SE Asian and Australian monsoons. This may be an interesting observation for further investigation by domain scientists.

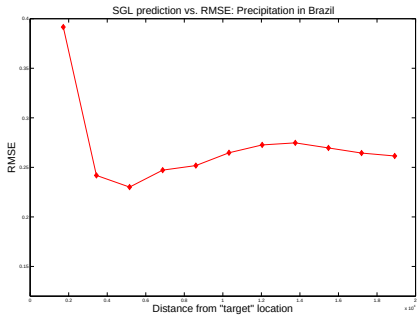
It should be noted that the dataset is a set of discrete samples from variables which vary continuously over space and time. This gives rise to 'sampling noise', which is manifested in some variables being selected by SGL, which might not have physical interpretations. Handling such data appropriately is a topic of future research.

6.5.3 Neighborhood Influence in Linear Prediction:

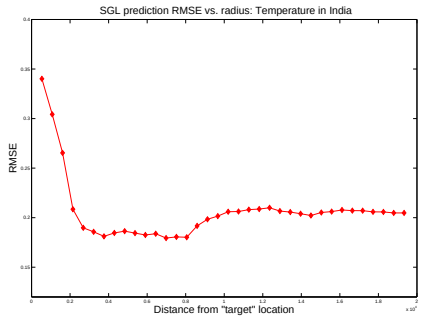
The previous discussion indicates that neighborhood sea locations play one of the most crucial roles in determining climate on land. We further investigate this fact through the following experiments.



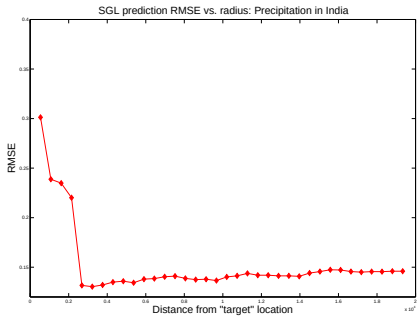
(a) Temperature prediction in Brazil.



(b) Precipitation prediction in Brazil.



(c) Temperature prediction in India.



(d) Precipitation prediction in India.

Figure 6.4: Comparison of SGL RMSE with distance (R beyond which all ocean locations are discarded. For small values of R, informative covariate locations are not included, and hence the predictive error is high. Adding more nearby informative locations decreases predictive error. Further addition of locations includes noisy covariates, leading to larger error in prediction.

We observe the RMSE on the test set from SGL regression by considering only those ocean variables which lie within a certain (geodesic) distance R from the target land region. We increase R from the 'smallest' distance, where only immediate neighborhood ocean locations of the target land region are considered, to the 'largest', when all locations on the earth are considered and note the change in RMSE of SGL prediction. Figs.6.4(a)-6.4(b) show the plots obtained for temperature and precipitation prediction in Brazil, while Figs.6.4(c)-6.4(d) show the same for India. The x-axis denotes the geodesic radius in kilometers from the target region within which all ocean covariates are considered, while disregarding all other ocean covariates outside this radius. The y-axis denotes the corresponding RMSE.

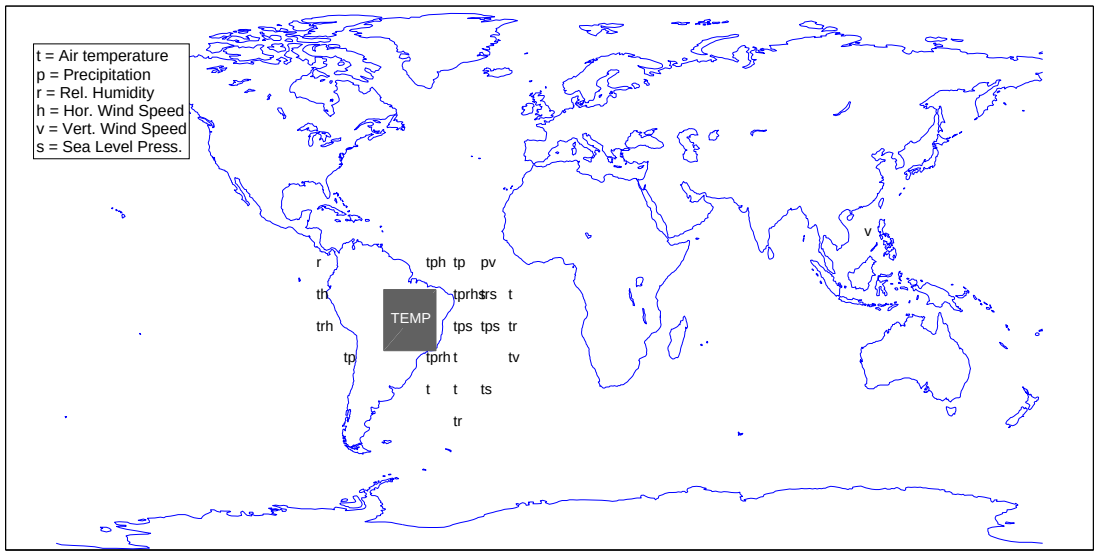


Figure 6.5: Temperature prediction in Brazil: SGL selected variables. All the plotted variables are selected in every single run of cross-validation.

The plots show that the least error in prediction is obtained when we include covariates in locations which are in the immediate neighborhood of the target variable. Omitting some of the locations leads to a sharp decrease in predictive power. This is consistent with our previous observation that SGL captures high proximity-dependence of the target variables. Covariates which are far away lead to a small increase in RMSE. It may be because most of these covariates are *irrelevant* to our prediction task and appear as "noise". However, the power of the SGL model lies in the fact that it can "filter" out this noise by having much smaller weight on some of these covariates and zero weight on others. The RMSE curve shows a number of 'dips', which might denote that there exist covariates with high predictive power at that distance, which, on being included, increase predictive accuracy of the model.

6.6 Variable Selection by SGL

As we noted earlier, the regularization parameters (α, λ) play a crucial role in variable selection. It is, therefore, noteworthy to study how variable selection changes with the change in the parameter values. For each covariate, we can compute and plot the coefficient value for a set

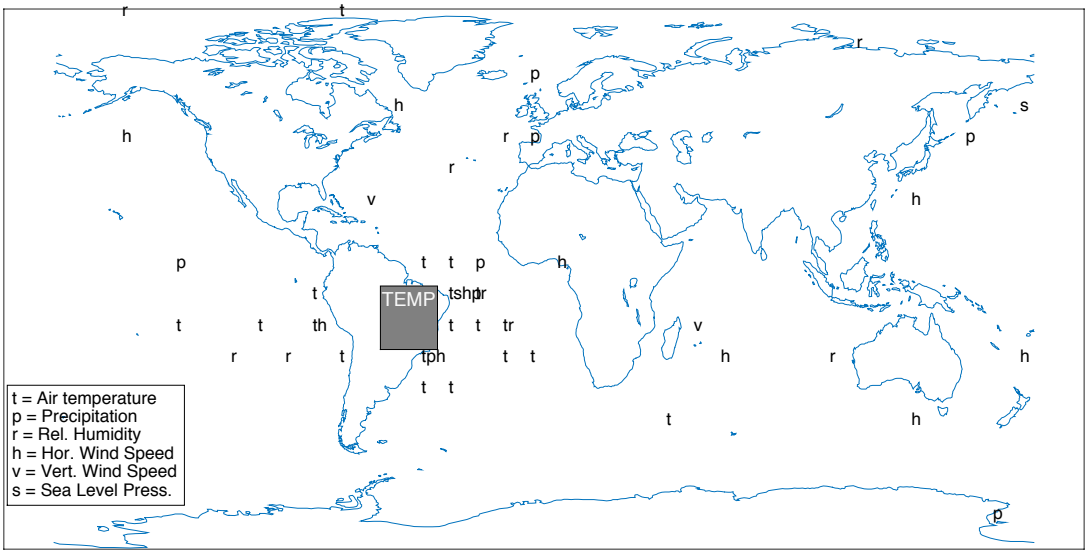


Figure 6.6: Temperature prediction in Brazil: LASSO selected variables. All the plotted variables are selected in every single run of cross-validation.

of chosen (α, λ). Thus, this plot, referred to as the *Regularization Path* of the SGL solutions [25], illustrates how the coefficient values change with change in penalty λ acts as a "tuning" parameter for the model. With higher penalties, we obtain a sparser model. However, it usually corresponds to a gain in RMSE. Most importantly, though, we obtain a quantitative view of the complexity of the model. In particular, the covariates which persist over considerably large ranges of λ and α are the most *robust* covariates in our regression task.

For the chosen training and test datasets, we compute the regularization path for temperature and precipitation predictions in Brazil and India. We fix $\alpha = 0.5$, so that $\lambda_1 = \lambda_2 = \frac{\lambda}{2}$. Figs.6.8(a) - 6.8(b) show the regularization paths for prediction in Brazil. The most 'stable' covariates, viz. temperature and precipitation in location(s) just off the coast of Brazil, have been earlier reported as among the most *relevant* covariates obtained through cross-validation on the training set.

The regularization paths for prediction in India are plotted in Figs.6.8(c) - 6.8(d). We observe that in this case too, the most stable covariates are among the *relevant* ones obtained through cross-validation. It is interesting to note that in all the plots, for low values in penalty, a mild increase in penalty dramatically changes the selected model. However, in higher ranges,

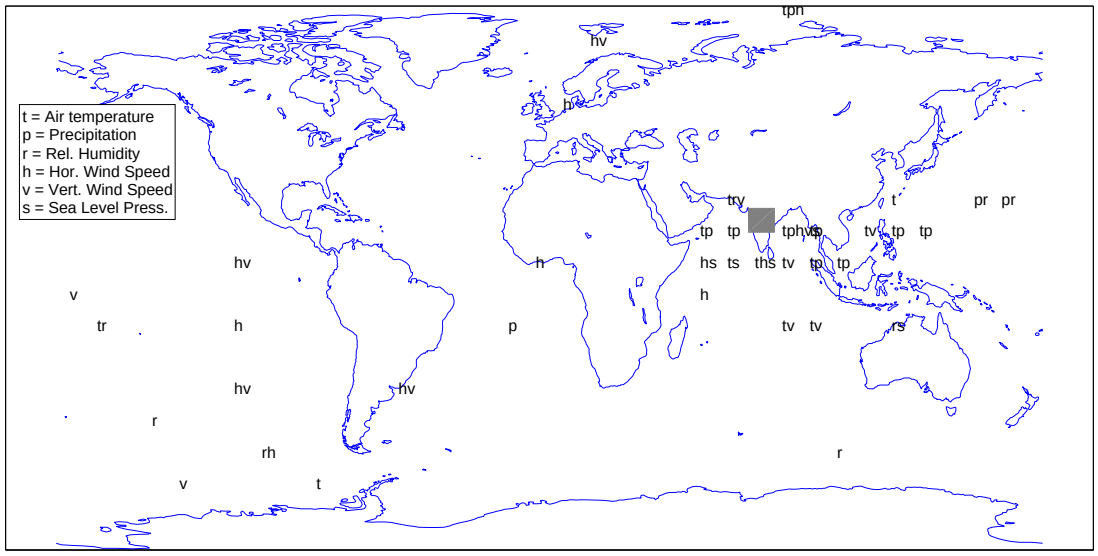


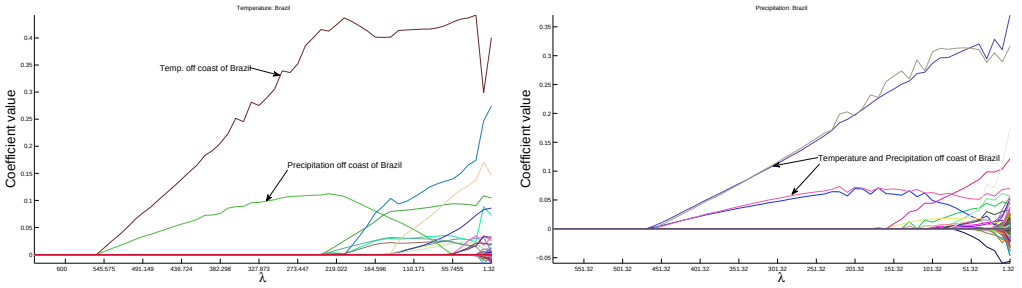
Figure 6.7: Temperature prediction in India: Variables chosen through cross-validation.

since the only covariates which survive are the relevant and stable ones, the change in model selection is more gradual.

6.7 Regularization Paths

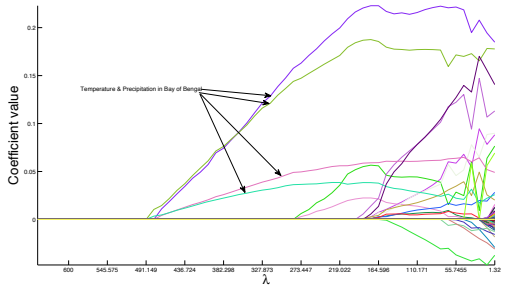
As we noted earlier, the regularization parameters (α, λ) play a crucial role in variable selection. It is, therefore, noteworthy to study how variable selection changes with the change in the parameter values. For each covariate, we can compute and plot the coefficient value for a set of chosen (α, λ). Thus, this plot, referred to as the *Regularization Path* of the SGL solutions [25], illustrates how the coefficient values change with change in penalty λ acts as a "tuning" parameter for the model. With higher penalties, we obtain a sparser model. However, it usually corresponds to a gain in RMSE. Most importantly, though, we obtain a quantitative view of the complexity of the model. In particular, the covariates which persist over considerably large ranges of λ and α are the most *robust* covariates in our regression task.

For the chosen training and test datasets, we compute the regularization path for temperature and precipitation predictions in Brazil and India. We fix $\alpha = 0.5$, so that $\lambda_1 = \lambda_2 = \frac{\lambda}{2}$. Figs.6.8(a) - 6.8(b) show the regularization paths for prediction in Brazil. The most 'stable'

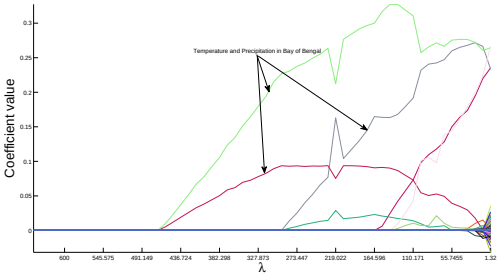


(a) Temp. prediction in Brazil

(b) Precip. prediction in Brazil



(c) Temp. prediction in India



(d) Precip. prediction in India

Figure 6.8: Regularization Paths for SGL on four use cases

covariates, viz. temperature and precipitation in location(s) just off the coast of Brazil, have been earlier reported as among the most *relevant* covariates obtained through cross-validation on the training set.

The regularization paths for prediction in India are plotted in Figs.6.8(c) - 6.8(d). We observe that in this case too, the most stable covariates are among the *relevant* ones obtained through cross-validation. It is interesting to note that in all the plots, for low values in penalty, a mild increase in penalty dramatically changes the selected model. However, in higher ranges, since the only covariates which survive are the relevant and stable ones, the change in model selection is more gradual.

6.8 Conclusion

Structured regression methods provide powerful tools for high dimensional data analysis problems encountered in climate science. In this chapter, we considered the task of predicting

climate variables over land regions using climate variables measured over oceans. We illustrated that the sparse group lasso (SGL) can encode sparsity arising from the natural grouping within covariates, using a hierarchical norm regularizer. Experiments prove improved prediction accuracy, and interpretable model selection by SGL compared to ordinary least squares. Further, feature selection was robust, and more interpretable than unstructured regularization using Lasso. Thus structured estimation methods hold enormous promise for applications to statistical modeling tasks in varied climate science problems.

Chapter 7

Understanding Dominant Factors for Precipitation over the Great Lakes Region

7.1 Motivation

Understanding climate change and its impacts on policy and infrastructure involves prediction of state of earth's climate under different forcing scenarios [120]. One of the most important variables of interest in modeling climate is precipitation, particularly at regional or local scales. Earth System Models (ESM) [121] that model the physics and dynamics of climate, are known to have deficiencies in modeling local precipitation [122, 123, 124]. This shortcoming is mainly due to the spatial resolution of the models, which is often too coarse to accurately model local and regional precipitation [123]. Therefore, although the physics of how precipitation occurs is well known, there exists a gap in understanding of the factors affecting precipitation over small scale regions on the globe.

Increasingly, statistical models are being considered to inform climate science research on factors which may affect precipitation [16]. The goal is to discover statistical dependencies between precipitation and covariates of interest, and then try to gain a mechanistic physical understanding of how the covariates affect precipitation. The covariates or predictors are often multi-scale climate variables and processes, which may manifest their effect with some temporal

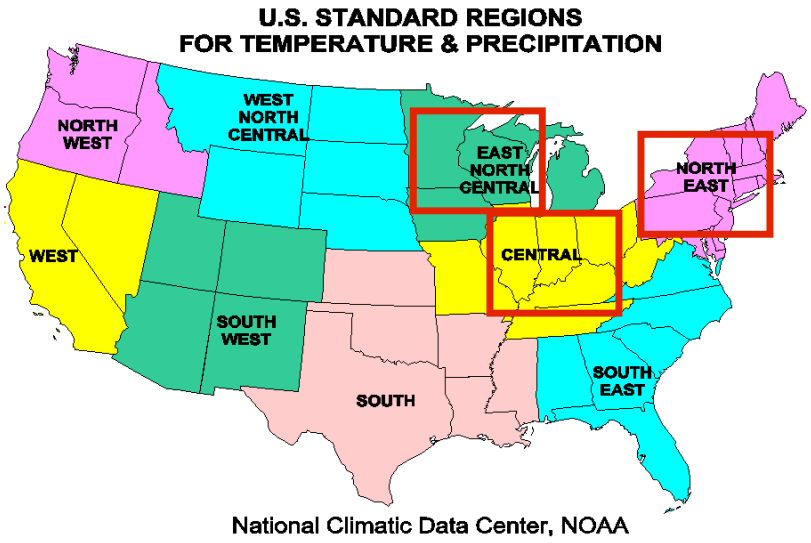
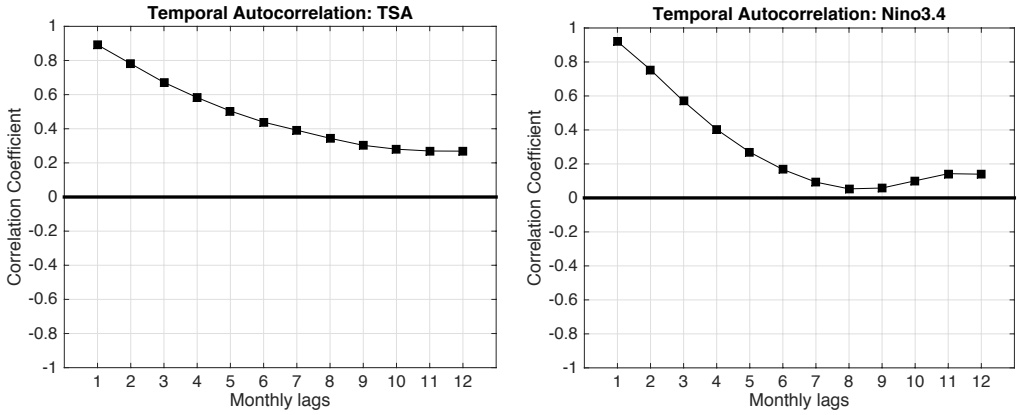


Figure 7.1: U.S. Standard Climatological Regions [2]. The Great Lakes consist of the three marked regions.

lags, or in conjunction with each other [125]. For a given region of interest, there is a plethora of possible influencing factors for precipitation, such as ocean oscillations, atmospheric variables, and long-term ocean-atmosphere coupled processes [38]. Therefore, it is of interest to the climate research and modeling community to understand the most influential factors in this pool of predictors, and derive climatological insights from such a discovery process.

In this work, we consider prediction of precipitation over the Great Lakes region of the US (Fig. 7.1), using predictor variables at multiple spatial scales with temporal lags. The predictors include atmospheric variables at local and regional scales, as well as multiple global climate indices [126] that capture climate processes and oscillations. We consider the December-January-February (DJF) or winter mean precipitation at a given weather station in the region as the response. The goal, therefore, is to understand the dominant factors for precipitation from among the large pool of possible predictors.

Sparse regression methods, such as LASSO [27], are useful in this scenario. Such methods allow simultaneous feature selection and regression, and are often supported by theoretical guarantees [30, 31]. LASSO has been found to perform well empirically in multiple other domains, and also provides fast solvers for efficient implementation [65]. However, often predictors have

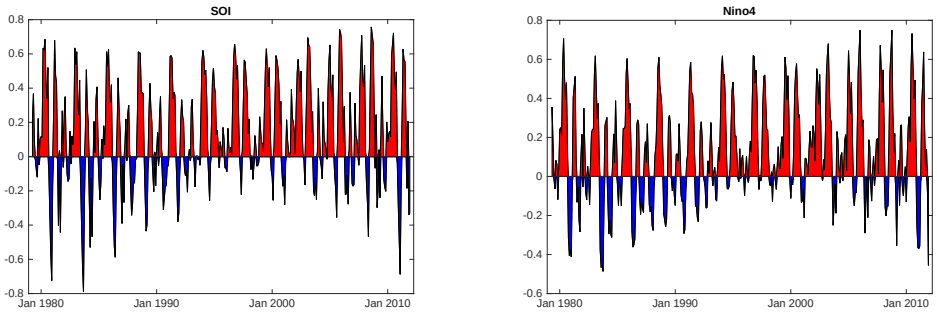


(a) TSA Index (b) Nino3.4 Index

Figure 7.2: Temporal Autocorrelations in climate indices. Some indices, such as TSA have significant correlations for upto 11 months.

temporal auto-correlations (Fig. 7.2), and since stations located in a region have geographical proximity, the data samples are also spatially correlated. Further, different climate indices related to the same climate phenomenon may be mutually correlated (Figs. 7.3(b) and 7.3(a)). In presence of such correlations, the set of features selected by LASSO may exhibit instability, and may include spurious predictors. In order to address this issue, we consider significance testing of selected set of predictors, to obtain stable and statistically significant covariates as dominant predictors. We use a random permutation test [127], to test the significance of each selected predictor, followed by composite analysis to gain a physical understanding of the effect of covariates on precipitation.

The rest of the chapter is arranged as follows. In Section 7.2, we overview the sparse regression methodology, and the random permutation testing framework. We describe the dataset used and pre-processing techniques in Section 7.3. In Section 7.4, we present experimental results, and discussions. Finally, we conclude in Section 7.5.



(a) Monthly SOI index

(b) Monthly Nino4 index

Figure 7.3: Climate Indices over Pacific which capture the El-Nino Southern Oscillation (ENSO)

7.2 Sparse Regression for Feature Selection

Sparse regression allows one to simultaneously conduct feature selection and regression, thus enabling selection of the most predictive set of features. We consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}\,, \tag{7.1}$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times p}$ are samples and $\boldsymbol{\beta}^*$ is a p-dimensional coefficient vector. The LASSO [27] method estimates a sparse $\hat{\boldsymbol{\beta}}$, by solving the following estimation problem:

$$\hat{\boldsymbol{\beta}} = \arg,\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\,, \tag{7.2}$$

where $\lambda > 0$ is a regularization parameter. In the context of discovering the dominant factors, often there is no prior bias on the sparsity imposed on the coefficients, although some of the covariates considered in the model may have strong correlations among each other, and temporal autocorrelation within itself over monthly or seasonal values. For example, consider the Nino4 index (Fig. 7.3(b)), which is computed from sea surface temperature, and the SOI index (Fig. 7.3(a)), which is derived from sea level pressure. Both indices carry information regarding the El-Nino Southern Oscillation (ENSO) [128], albeit from different climate variables. Hence they exhibit a high negative correlation (about -0.6). Some of the indices also show high temporal autocorrelations (Fig 7.2).

In presence of such correlations in covariates and samples, the set of features selected by LASSO often have instability [129]. Further, for finite samples, there is a non-zero probability

that for a given training set and a chosen penalty parameter LASSO selects a non-zero coefficient for a non-informative predictor by random chance. Therefore, we require a significance testing method to test each non-zero coefficient estimated by LASSO on training data, and compute a p-value for significance of each feature.

Testing significance of covariates has been considered in various problems of applied statistics, and the most commonly used testing methodology is random permutation test [130, 131, 132]. Such a test is a nonparametric hypothesis testing framework, which measures the significance of every non-zero coefficient value by constructing a random distribution over the coefficient using random permutations of the data. We adopted a variation of the methodology developed in [127] that we discuss next.

Permutation Test

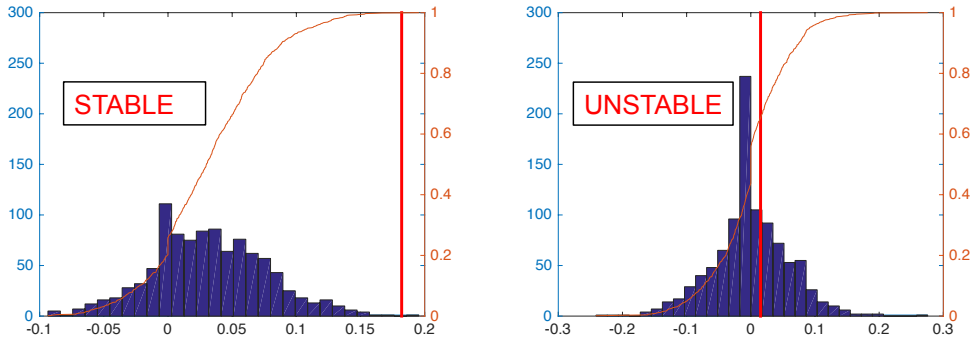
We fix λ at a particular value. On the training data, we first compute the LASSO estimate $\hat{\boldsymbol{\beta}}$ by solving (7.2). Next, keeping $\mathbf{X}$ constant, we randomly permute the response $\mathbf{y}$ to obtain a vector $\tilde{\mathbf{y}}$. The random permutation of the response destroys any statistical relationship existing between the covariates in $\mathbf{X}$ and the response $\tilde{\mathbf{y}}$. Thereafter, we run LASSO with $\mathbf{X}$ and $\tilde{\mathbf{y}}$ in order to obtain a random coefficient vector $\tilde{\boldsymbol{\beta}}$, which represents random causal relationships between the covariates and the response. Executing this strategy multiple ($\nu \geq 1000$) times, for the i-th non-zero coefficient in $\hat{\boldsymbol{\beta}}$, we compute the probability that a random value $|\tilde{\boldsymbol{\beta}}_i|$ exceeds the estimated value $|\hat{\boldsymbol{\beta}}_i|$ given by

$$p_i = \frac{count(|\tilde{\boldsymbol{\beta}}_i| \geq |\hat{\boldsymbol{\beta}}_i|)}{\nu + 1} \,. \tag{7.3}$$

It represents the p-value associated with the corresponding coefficient $\hat{\boldsymbol{\beta}}_i$.

Fig. 7.4 illustrates the results of permutation test on two coefficients i and j. the coefficient i is stable since it lies at the tail of the empirical distribution and thus has low p-value (< 0.05) so that we can reject the null hypothesis that the estimated value occurred due to random chance. However, for coefficient j, the estimated value lies near the mode, and hence obtains a high p-value.

In Fig. 7.5, we plot the stable features that are selected using LASSO and permutation test in the training set. Evidently, increasing the regularization parameter λ in LASSO leads to pruning and we obtain a smaller set of stable parameters. However, note that the permutation



(a) Stable coefficient i (b) Unstable coefficient j

Figure 7.4: Behavior of Stable and Unstable variables during random permutation test. The histogram represents an empirical approximation of the distribution of the coefficient value under the null hypothesis that $\mathbf{y}$ is exchangeable. A low p-value (a) shows that the estimated value lies to the tail of the distribution.

test for each value of λ is independent, and therefore the pruning exhibited in Fig. 7.5 is a sign of stability of the selected features, rather than an artifact of the LASSO solution.

7.3 Dataset

We compiled datasets from two sources: (1) United States Historical Climatological Network (USHCN) [133], and (2) North American Regional Reanalysis (NARR) [134]. We considered the following 8 states as consisting the Great Lakes region of the USA: (i) Minnesota (MN), (ii) Wisconsin (WI), (iii) Illinois (IL), (iv) Indiana (IN), (v) Michigan (MI), (vi) Ohio (OH), (vii) Pennsylvania (PA) and (viii) New York (NY). We further aggregated the states to lie in one of the three climatological regions in the Great Lakes (Fig. 7.1). For each state, station level data for daily maximum/minimum temperature, and precipitation was available for each station in the state. We considered the average winter (DJF) precipitation for each station as a response. Therefore, for every region, we had winter precipitation data for stations for 1979-2011.

The covariates consisted of local, regional and global climate variables (listed in Table 7.1). We considered local and regional surface temperature and pressure (SLP) and convective available potential energy (CAPE) over winter (DJF) and autumn (SON) as covariates. For each

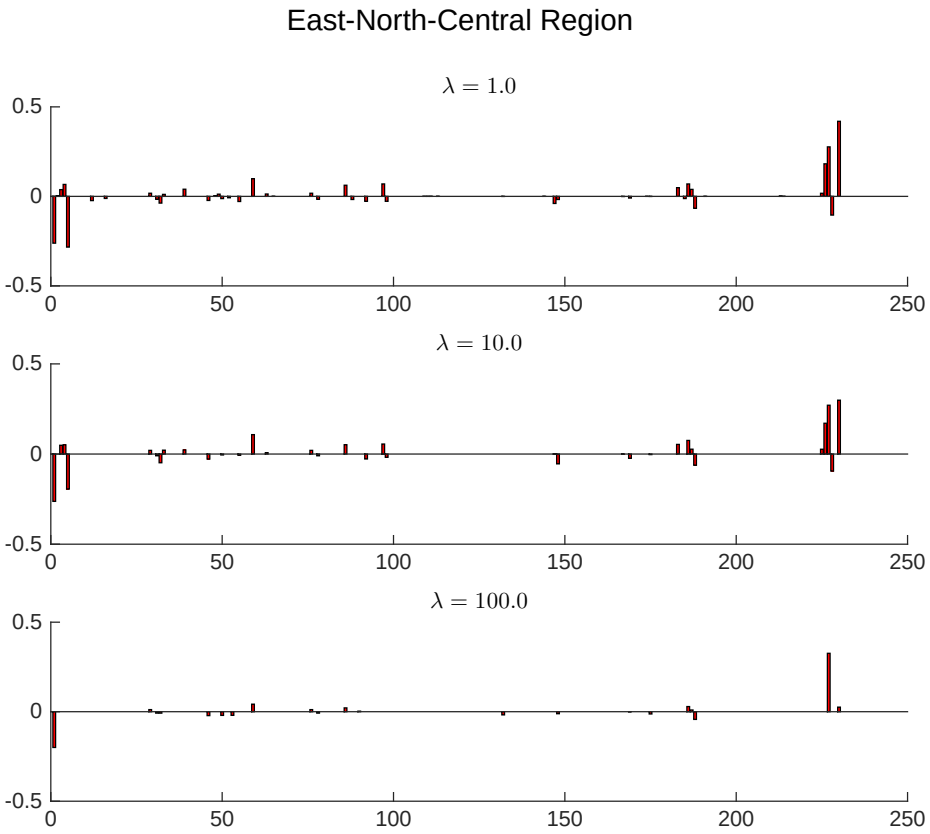


Figure 7.5: Stability of dominant factors at different penalization values. At higher penalization values, the set of coefficients is pruned, but no additional coefficients are introduced into the set.

Table 7.1: Covariates for Precipitation prediction

Type	Variables
Atmospheric (Station level)	Winter Minimum Temperature (DJF_Tmin), Winter Maximum Temperature (DJF_Tmax), Autumn Minimum Temperature (SON_Tmin), Autumn Maximum Temperature (SON_Tmax), Sea Level Pressure (SLP), Convective Available Potential Energy (CAPE), Air Temperature at 500mb (AIR_500)
Atmospheric (Regional averages)	Regional Average Winter Minimum Temperature (DJF_TRegmin), Regional Average Winter Maximum Temperature (DJF_TRegmax), Regional Average Autumn Minimum Temperature (SON_TRegmin), Regional Average Autumn Maximum Temperature (SON_TRegmax), Regional Average Sea Level Pressure (SLPReg), Regional Average Convective Available Potential Energy (CAPEReg), Regional Average Air Temperature at 500mb (Reg_AIR_500)
Large-Scale Climate Indices	North Atlantic Oscillation (NAO), East Atlantic Pattern (EA), West Pacific Pattern (WP), East Pacific/North Pacific Pattern (EPNP), Pacific/North American Pattern (PNA), East Atlantic/West Russia Pattern (EAWR),Scandinavia Pattern (SCA), Tropical/Northern Hemisphere Pattern (TNH), Polar/Eurasia Pattern (POL), Pacific Transition Pattern (PT), Nino 1+2, Nino 3, Nino 3.4, Nino 4, Southern Oscillation Index (SOI), Pacific Decadal Oscillation (PDO), Northern Pacific Oscillation (NP), Tropical/Northern Atlantic Index (TNA), Tropical/Southern Atlantic Index (TSA), Western Hemisphere Warm Pool (WHWP)

global climate index, we considered all 12 preceding values (from Jan to Dec. of a year) as covariates. We discarded the lower and upper one percentile of the precipitation data since these correspond to very low and very high precipitation, and therefore are "extreme events" [16], which often have very different mechanisms than normal precipitation [135, 136]. In total, the dataset had 2200 samples over 32 years, where we discarded samples which contained missing values. We divided the data into two sets. The first, comprising of 22 years' data, was used for finding dominant factors. The second set, with the remaining 10 years' data, was used to test

predictive performance. We discuss these in detail in the next section.

7.4 Results and Discussion

We used a randomly selected 22 years' data for obtaining the dominant features. Further, we conducted leave-one-out crossvalidation on the remaining 10 years' data to test the predictive performance of the dominant predictors.

7.4.1 Predictive Performance

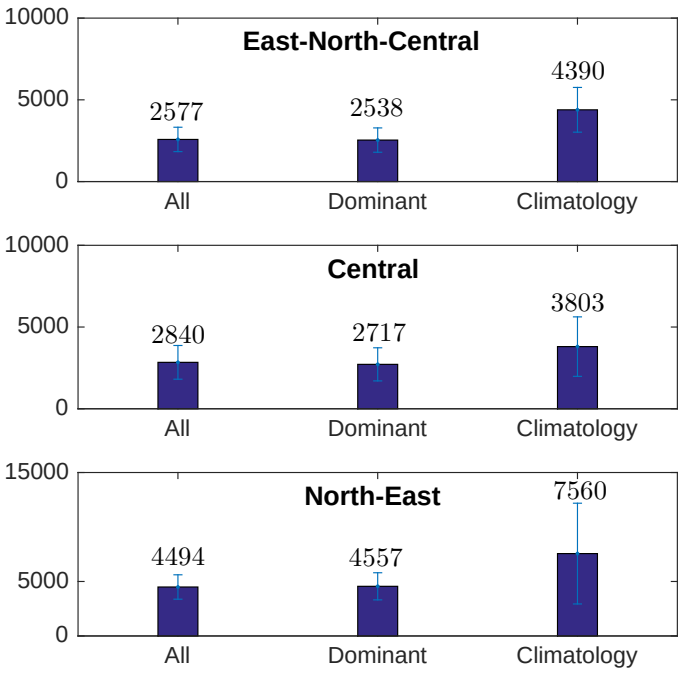


Figure 7.6: Mean Square Error on precipitation measured at hundredths of an inch of Ordinary Least Squares regression using only dominant factors and using all covariates. The prediction errors from long-term climatology is also plotted. The error bars denote one standard deviation.

It is important to assess the predictive performance of the dominant factors found by the

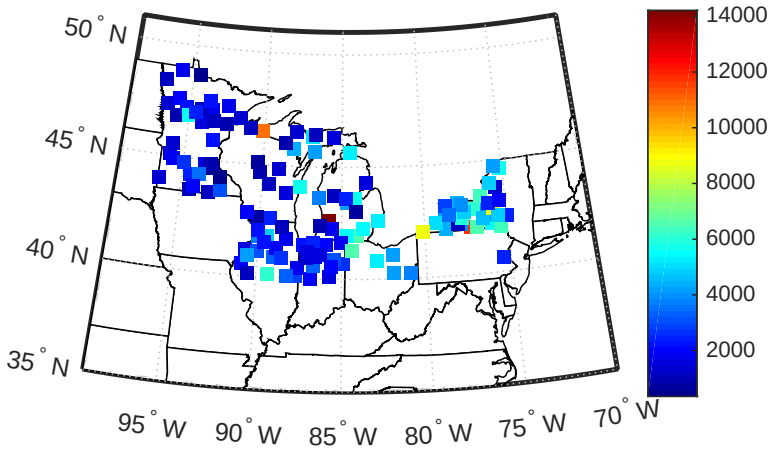


Figure 7.7: Geographical Spread of Errors (MSE) over the region. The North-East has higher errors than inland regions.

proposed method against the climatology of each region. The climatology denotes the long-term average of precipitation over the region. Predictive covariates need to show improvement upon the prediction from long-term climatology, in order to be considered for further hypothesis generation on the mechanism of precipitation.

We conducted leave-one-year-out cross-validation on held out test set described earlier. Fig. 7.6 shows the mean square error (MSE) from ordinary least squares regression using dominant factors (less than 25 factors in each region) vs. the entire pool of 232 predictors. The performance is identical (2-sample t-test p-value more than 0.8 on all three cases), and much better than simply predicting the climatological mean. This illustrates that the dominant predictors carry almost all predictive information available in the set of covariates.

Further, for each station, we computed the MSE in the test set during crossvalidation. In Fig. 7.7, we have plotted MSE at each geographic location of the stations. MSE in the inland locations (Central and East-North-Central region) are lower than in the North-East region. Higher MSE in the North-East is understandable due to the complex processes which affect variation of precipitation in this region. The north pacific jet stream and the Lake Effect often causes large

variation, along with influences of winds from Atlantic, since the area is near the coast.

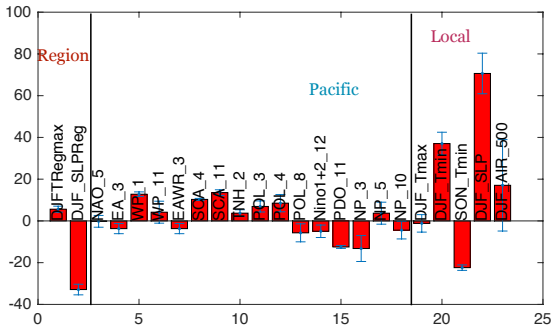
7.4.2 Dominant Factors

For each climatological region, we obtained a subset of features as the dominant factors, which are plotted in Fig. 7.8. We fix λ at a value which provides low prediction mean square error (MSE) on a small validation set. In Fig. 7.8, for each selected factor, we also plot the mean and standard deviation (as error bars) of the coefficient obtained during the leave-one-out crossvalidation. It is evident from the thin error bars that each factor gets assigned stable coefficients. Some interesting patterns emerge from these figures. Surface air temperature in winter plays a prominent role in precipitation during the winter season. It is well known that high snowfall years typically experience lower than normal minimum temperature. Moreover this effect is more pronounced in inland regions (East-North-Central and Central). However, note that since heavy precipitation may itself lower the surface temperature, the relationship may depict correlation rather than causation.

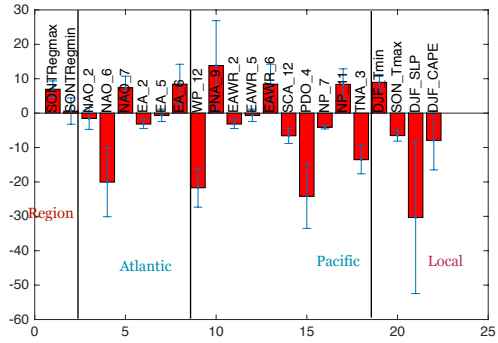
Sea level pressure (SLP), which is a dominant factor in the ENC and Central regions, has great influence on the surface level winds that carry moisture from the Pacific across the continent to the Great Lakes. Lower SLP over the region is often associated with higher moisture flow and thus higher precipitation. However, since variations of SLP is a surface phenomenon, it is more noisy as a predictor in seasonal scales than higher atmospheric variables. Therefore, we obtain higher variance in the weights for SLP over crossvalidation runs.

The North-East region behaves differently in that only a single local atmospheric variable is selected as dominant factor. This may be indicative of the fact that the North-East region, due to its proximity to the ocean, is influenced heavily by oceanic effects. As noted earlier, it is known that there are multiple factors for variation of precipitation over this region. For example, the Lake Effect [137] has substantial impact on snowfall during winter, which is influenced by the Pacific jet stream. Due to this phenomenon, often the region experiences very heavy snowfall over only a few days or hours.

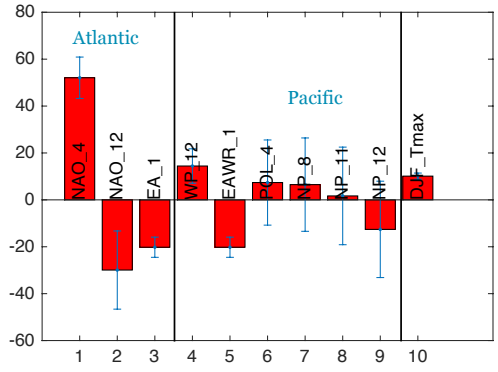
Atlantic and Pacific influences are prominent across the entire Great Lakes region. Moreover, comparison of the three panels of Fig. 7.8 shows that Atlantic influences become more prominent on the eastern part of the Great Lakes, while most stable indices in the ENC region are computed over Pacific. Particularly, consider the dominant factors of precipitation over the ENC region. The dominant climate indices are mainly EA (East Atlantic Pattern), WP (West



(a) East-North-Central

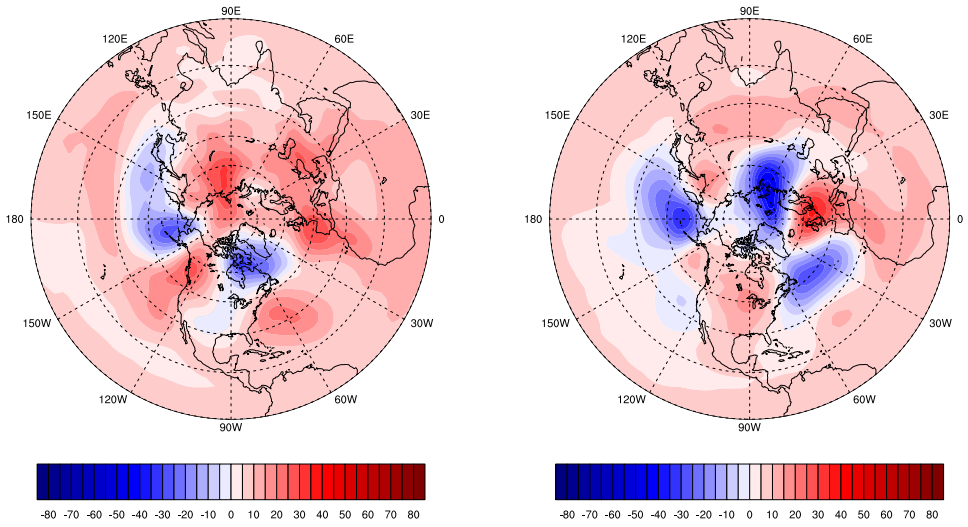


(b) Central



(c) North-East

Figure 7.8: Dominant factors for precipitation in each region. The standard abbreviation for each index has been used, along with the month represented as a number. Influences from Atlantic and Pacific are evident in all three regions, mainly from tropical and east pacific, and north atlantic. Multiple summer index values are deemed significant. Further local atmospheric influences are deemed more predictive for inland regions, while oceanic indices are the sole dominant factors in the maritime region.



(a) High precipitation

(b) Low precipitation

Figure 7.9: Average 700mb Geopotential height anomalies in December over (a) 10 highest precipitation years, and (b) 10 lowest precipitation years. Note the strong negative anomaly (low pressure) in the years with high precipitation which causes increased moisture to flow in from the Pacific.

Pacific Pattern), SCA (Scandinavian Pattern), TNH (Tropical/Northern Hemisphere Pattern), POL (Polar/Eurasia Pattern), PDO (Pacific Decadal Oscillation) and NP (Northern Pacific Oscillation). All of these indices are computed from or have high correlation with 700mb – 500 mb geopotential height anomalies. Therefore, we construct composites for geopotential height anomalies in order to further investigate the processes leading to precip variations across the ENC region.

7.4.3 Composites over Geopotential Height Anomalies

In Fig. 7.9, we plot average December 700mb geopotential height anomalies over the northern hemisphere, where the average is taken over the 10 highest and 10 lowest precipitation years. Fig. 7.9(a) shows a strong negative anomaly over Canada and north-central U.S. denoting existence of a low pressure system. The strong low pressure system is conducive for increased

wind flow from northern Pacific, which picks up moisture from the Pacific ocean and thus favors higher moisture content in the air. In the presence of colder temperatures over much of the region, this may lead to increased precipitation.

In stark contrast, Fig. 7.9(b) illustrates that a *positive* anomaly exists over the entire U.S. for seasons with low precipitation. Such anomalies are associated with higher than average pressure system over the region, and may adversely affect precipitation in two ways. First, since the Pacific has negative anomalies (Fig. 7.9(b)), the system is not conducive for wind flowing into the continent from the Pacific. Thus it leads to less moisture flow into the region. Second, the positive anomalies at higher levels (700mb) may also lead to down drafts from the upper atmosphere, thus decreasing convective precipitation.

The two panels in Fig. 7.9 represent *typical* patterns for geopotential height anomalies over the northern hemisphere for high and low precipitation seasons. The climate indices discussed previously seem to capture these typical patterns and provide predictive information using such patterns. That such typical patterns have predictive information is clear from Fig. 7.6, since otherwise the performance of the predictive model would not be better than the climatology of the region. Further such patterns are often persistent over months leading to winter. Fig. 7.10 illustrates the geopotential height anomalies averaged over 10 highest precipitation years over ENC region. The low pressure region moves East from over Pacific in September to over U.S. in December, which is consistent with the movement of the Westerlies.

In Fig. 7.11, we plot the anomalies in geopotential height, similarly, for averaged over the 10 lowest precipitation years in the ENC region. Note the high pressure system which moves from Siberia in September to North America across the Pacific Ocean. The high pressure system obstructs moisture flow into the Great Lakes, and also causes downdraft from upper atmosphere, thus reducing convection.

7.5 Conclusions

In this chapter, we proposed a method for discovery of dominant factors for precipitation over the Great Lakes region using a sparse regression method, in conjunction with permutation test for significance. Dominant factors discovered through this process showed high predictive power and produced lower error than obtained from climatology. Further, composite analysis

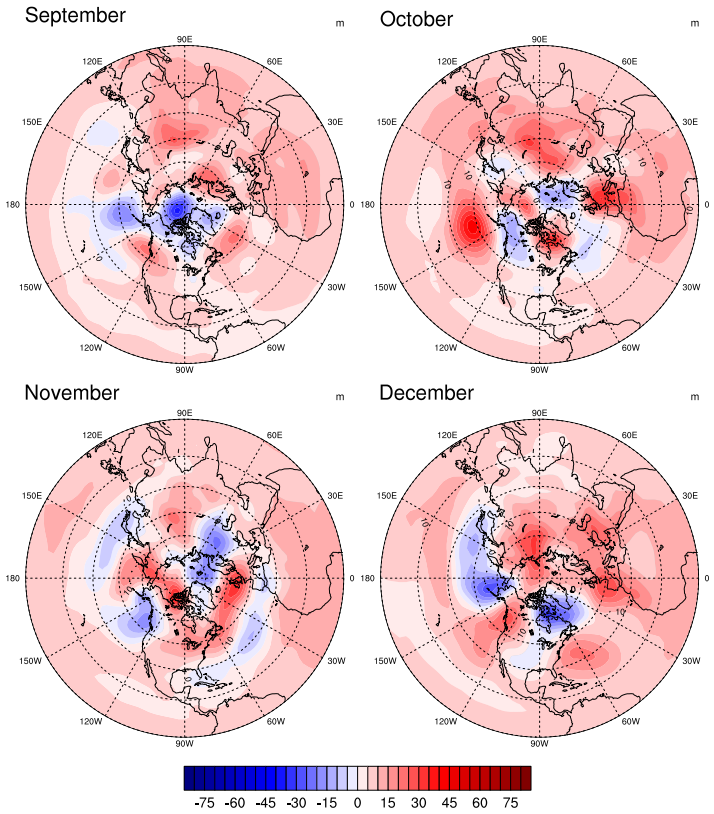


Figure 7.10: Geopotential height anomalies averaged over 10 highest precipitation years over ENC region in months leading to winter. The low pressure region shifts from Pacific to over the U.S. over the Fall months along the westerlies.

of some of the discovered factors shows that atmospheric patterns persisting over entire seasons may affect precipitation over the region, and is consistent with understanding from climate science. Thus, the proposed method may be useful for deriving hypotheses over how stable atmospheric patterns, such as variations in geopotential heights, may produce scenarios which influence wind and moisture flow, and thus precipitation. In general, the method will be useful for constructing such hypothesis in various statistical modeling scenarios in climate, which can then be further investigated for statistical and physical significance.

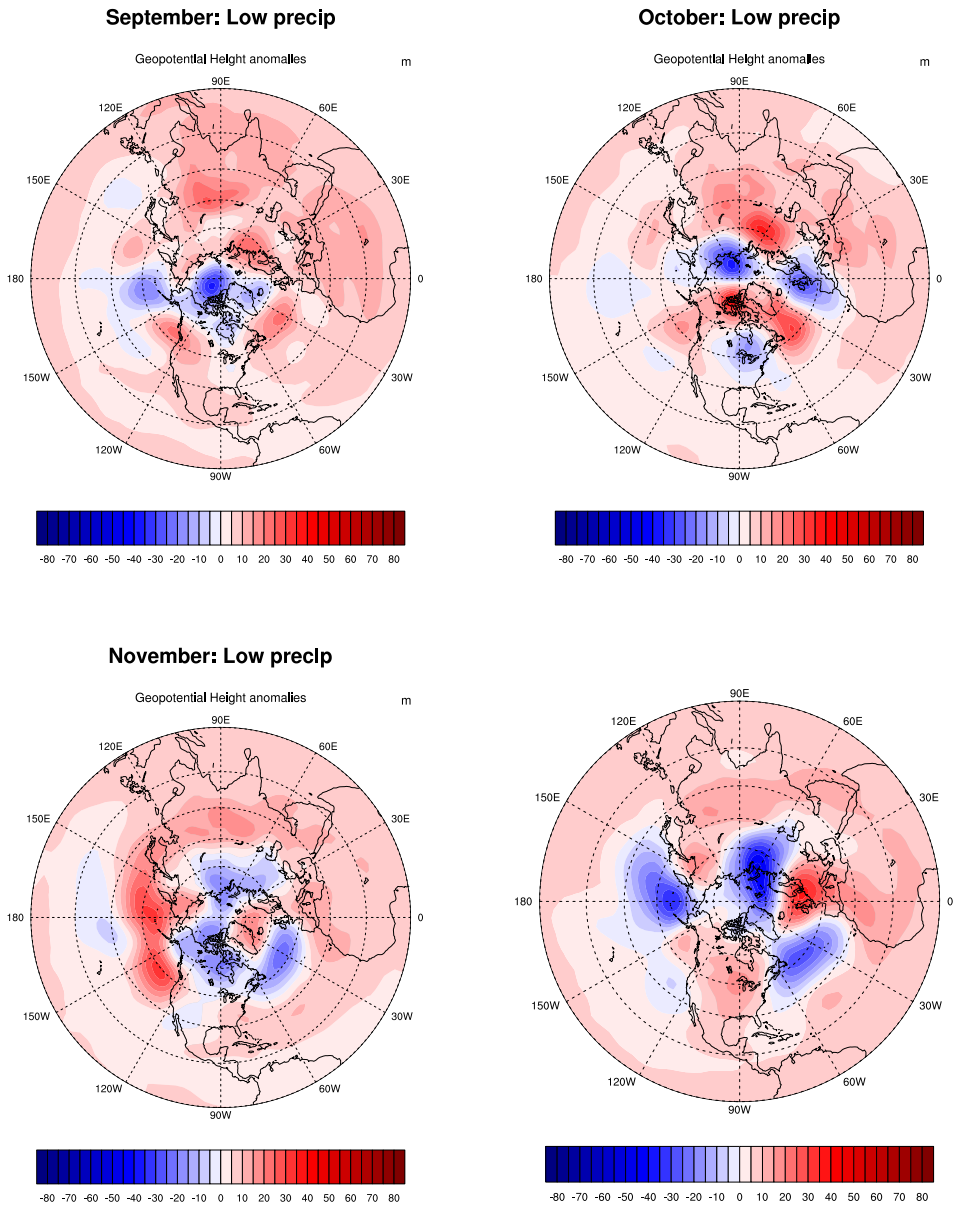


Figure 7.11: Average Geopotential height anomalies over the 10 lowest precipitation years in ENC region for the months leading upto winter. Note the high pressure system over Siberia moves across the Pacific into North America. The Polar Pattern (POL) is a dominant factor for the ENC region and is closely related to this pressure system.

Chapter 8

Conclusions

In this thesis, we have presented our research on high dimensional models that advances understanding in two directions: theoretical advancements in regularized regression with structured regularizers, and applications of structured regression to statistical modeling problems in climate science.

In Chapter 3, we considered structured regression with hierarchical tree-structured norm regularizers. We illustrated that the hierarchical norm is a decomposable norm regularizer, which includes the popular group-lasso norm as a special case. We proved that estimation with hierarchical regularization is statistically consistent, and provided rates of consistency for the hierarchical Lasso, and the special case, Sparse Group Lasso.

In Chapter 4, we considered the Dantzig Selector for sparse linear regression, and generalized it to incorporate any norm regularizer. We illustrated that analysis of the Generalized Dantzig Selector requires two conditions, the restricted eigenvalue condition and an upper bound on the Gaussian width of the unit norm ball. Further, we considered the k-support norm, which enables selection of latent groups of covariates. We proved the first result on statistical consistency of estimation with the k-support norm using GDS. We also showed that for LASSO and ridge regression, which are special cases of k-support norm, the rates match existing results.

Real world scenarios often involve estimation with covariates corrupted with noise. In Chapter 5, we studied GDS estimation with noisy covariates. We showed that the analysis tools for the noiseless setting fail to prove statistical consistency in presence of noise. However, with an appropriate correction in the covariance matrix, GDS is provable consistent.

In the second part of the thesis, we provided two applications of high dimensional modeling

in climate science. First, we considered the modeling of interactions between land atmospheric variables and ocean variables in Chapter 6. The spatial nature of covariates lend itself to the grouping structure encoded by Sparse Group Lasso (SGL). We illustrated that sparse regression methods provide better prediction accuracy than unregularized regression. Further, structured regression with SGL provided more interpretable results than LASSO, while maintaining similar predictive performance, which is further validated by the regularization paths for SGL.

Lastly, in Chapter 7, we applied the sparse regression methodology to the task of discovering dominant factors for winter precipitation over the Great Lakes region. Due to temporal and spatial correlation existing within the climate variables, we required a significance testing framework using random permutation tests for selecting stable sets of predictors over a climatological region. Crossvalidation tests illustrated that dominant predictors have significant predictive information, and have stable regression coefficients. Further analysis of some of the selected predictors with climate composites showed that the predictors are related to atmospheric patterns that have known relationships to North American precipitation. Thus the feature selection methodology provides interpretable results, and looks promising for further exploration.

Outlook

Application of machine learning to statistical modeling problems in climate science has the possibility of opening new doors in climate data analysis. An example is extending the framework proposed in Chapter 7 to generate physical hypotheses on the mechanism(s) of precipitation, particularly in various climatological regions across the globe. However, there exist three fundamental limitations to the current methods in the context of climate data.

First, precipitation and other climate variables rarely follow a Gaussian distribution. For example, Fig. 8.1 illustrates the empirical precipitation distribution over the East-North-Central region over the Great Lakes. There are multiple seasons and stations where very low precipitation was recorded. On the other extreme, the distribution has a heavy tail in the positive direction. Therefore, one needs transformation of the data to approximate gaussianity such that the theoretical guarantees of the regression methods continue to hold.

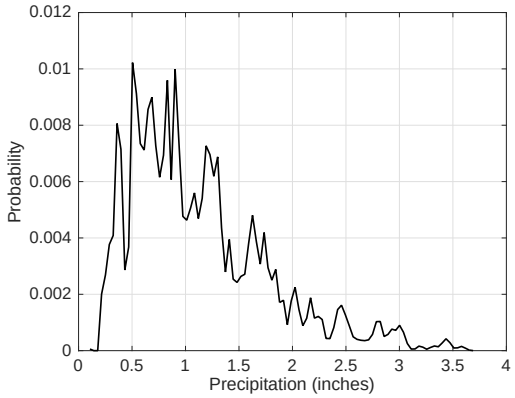


Figure 8.1: Empirical probability density function of average winter (DJF) precipitation (in inches) over the East-North-Central region.

Second, existing optimization of LASSO and other sparse regression methods mainly utilize proximal gradient descent or coordinate descent type algorithms, which rely on well-conditioned design matrix $\mathbf{X}$ for stable estimates. In presence of correlated covariates, the estimates are often unstable. However, climate indices often have high temporal autocorrelation, and are also mutually correlated. In presence of such multi-collinearity, designing efficient algorithms, as well as proving theoretical results are challenging.

Third, theoretical understanding of permutation tests for significance is, unfortunately, fairly limited. In presence of correlation within samples, such as spatial correlation between precipitation at nearby locations, it is unclear if and when permutation test fails to provide stable estimates. Understanding the conditions under which permutation tests may fail and designing tests for such conditions in climate datasets will greatly strengthen the results obtained in this domain.

References

[1] K. Steinhaeuser, N. Chawla, and A. Ganguly. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. *Statistical Analysis and Data Mining*, 4(5):497–511, 2011.

[2] Thomas Karl and Walter James Koss. *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*. National Climatic Data Center, 1984.

[3] Shaobing Chen and David Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

[4] Joel Tropp, Anna C Gilbert, et al. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[5] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[6] P Baldi and S Bruna. *Bioinformatics - The Machine Learning Approach*. MIT Press, 1998.

[7] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[8] Gavin C Cawley and Nicola LC Talbot. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348–2355, 2006.

[9] MK Stephen Yeung, Jesper Tegnér, and James J Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, 2002.

[10] Nicholas M Ball and Robert J Brunner. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, 2010.

[11] Alberto HF Laender, Berthier A Ribeiro-Neto, Altigran S da Silva, and Juliana S Teixeira. A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2):84–93, 2002.

[12] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.

[13] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001.

[14] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[15] Justin T Schoof and SC Pryor. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of Climatology*, 21(7):773–790, 2001.

[16] Debasish Das. *Bayesian sparse regression with application to data-driven understanding of climate*. PhD thesis, TEMPLE UNIVERSITY, 2015.

[17] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399, 2011.

[18] Christos Davatzikos, Kosha Ruparel, Yong Fan, DG Shen, M Acharyya, JW Loughead, RC Gur, and Daniel D Langleben. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, 28(3):663–668, 2005.

[19] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys & Tutorials, IEEE*, 12(2):159–170, 2010.

[20] JB Predd, SB Kulkarni, and HV Poor. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23:56–69, 2006.

[21] Debasish Das, Auroop R Ganguly, and Zoran Obradovic. A sparse bayesian model for dependence analysis of extremes: Climate applications. In *ICML 2013 Workshop on Inferning: Interactions between Inference and Learning*, 2013.

[22] Huahua Wang, Arindam Banerjee, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. Large scale distributed sparse precision estimation. In *Advances in Neural Information Processing Systems*, pages 584 – 592, 2013.

[23] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[24] Alex J Cannon. Negative ridge regression parameters for improving the covariance structure of multivariate linear downscaling models. *International Journal of Climatology*, 29(5):761–769, 2009.

[25] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. of Stats.*, 32:407–499, 2002.

[26] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning; Data mining, Inference and Prediction*. Springer Verlag, New York, 2001.

[27] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society Series B*, 68 (1):49–67, 2006.

[29] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

[30] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[31] Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems 27*, pages 1556–1564. 2014.

[32] Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *Signal Processing, IEEE Transactions on*, 59(9):4183–4198, 2011.

[33] Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop R Ganguly. Sparse group lasso: Consistency and climate applications. In *SDM*, pages 47–58. SIAM, 2012.

[34] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, December 2007.

[35] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[36] Intergovernmental Panel on Climate Change. Fifth assessment report, 2013.

[37] E. J. Plate. Flood risk and flood management. *Journal of Hydrology*, 267(1):2–11, 2002.

[38] J. Cunderlik and S. Simonovic. Inverse flood risk modelling under changing climatic conditions. *Hydrological processes*, 21(5):563–577, 2007.

[39] C. Castro and G. Pielke, R.and Leoncini. Dynamical downscaling: Assessment of value retained and added using the regional atmospheric modeling system (rams). *Journal of Geophysical Research: Atmospheres*, 110, 2005.

[40] L. O. Mearns, W. Gutowski, R. Jones, R. Leung, S. McGinnis, A. Nunes, and Y. Qian. A regional climate change assessment program for north america. *Eos, Transactions American Geophysical Union*, 90(36):311, 2009.

[41] R. L. L Wilby, T. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, 34(11):2995–3008, 1998.

[42] R. L. Wilby, S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns. Guidelines for use of climate scenarios developed from statistical downscaling methods. *IPCC Task Group on Data and Scenario Support for Impact and Climate Analysis (TGICA)*, 2004.

[43] P. Willems and M. Vrac. Statistical precipitation downscaling for small-scale hydrological impact investigations of climate change. *Journal of Hydrology*, 402(3):193–205, 2011.

[44] B. Hewitson and R. Crane. Consensus between GCM climate change projections with empirical downscaling: precipitation downscaling over South Africa. *International Journal of Climatology*, 26(10):1315–1337, 2006.

[45] Q. Schiermeier. The real holes in climate science. *Nature*, 463(7279):284–287, 2010.

[46] Timothy DelSole. A bayesian framework for multimodel regression. *Journal of climate*, 20(12):2810–2826, 2007.

[47] João Corte-Real, Xuebin Zhang, and Xiaolan Wang. Downscaling gcm information to regional scales: a non-parametric multivariate regression approach. *Climate Dynamics*, 11(7):413–424, 1995.

[48] Subimal Ghosh and PP Mujumdar. Statistical downscaling of gcm simulations to streamflow using relevance vector machine. *Advances in Water Resources*, 31(1):132–146, 2008.

[49] Robert L Wilby and TML Wigley. Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4):530–548, 1997.

[50] Soumyadeep Chatterjee, Sheng Chen, and Arindam Banerjee. Generalized dantzig selector: Application to the k-support norm. In *Advances in Neural Information Processing Systems*, pages 1934–1942, 2014.

[51] Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, et al. The ncep/ncar 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471, 1996.

[52] Ralph A Willoughby. Solutions of ill-posed problems (an tikhonov and vy arsenin). *SIAM Review*, 21(2):266–267, 1979.

[53] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM Journal on Matrix Analysis and Applications*, 21(1):185–194, 1999.

[54] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[55] George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129, 1994.

[56] Shelley Derksen and HJ Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45(2):265–282, 1992.

[57] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

[58] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[59] Peng Zhao and Bin Yu. On model selection consistency of lasso. *JMLR*, 7:2541–2563, 2006.

[60] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34 3:1436–1462, 2006.

[61] Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

[62] Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, pages 352–360, 2011.

[63] Sofia Mosci, Silvia Villa, Alessandro Verri, and Lorenzo Rosasco. A primal-dual algorithm for group sparse regularization with overlapping groups. In *Advances in Neural Information Processing Systems*, pages 2604–2612, 2010.

[64] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Preprint*, 2010.

[65] Jun Liu and Jieping Ye. Moreau-yosida regularization for grouped tree structure learning. In *NIPS*. 2010.

[66] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*.

[67] Han Liu, Jian Zhang, Xiaoye Jiang, and Jun Liu. The group dantzig selector. In Yee Whye Teh and D. Mike Titterington, editors, *AISTATS*, volume 9 of *JMLR Proceedings*, pages 461–468. JMLR.org, 2010.

[68] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[69] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[70] Yehoram Gordon. *On Milman's inequality and random subspaces which escape through a mesh in n*. Springer, 1988.

[71] Roman Vershynin. Estimation in high dimensions: a geometric perspective. *arXiv:1405.5103*, 2014.

[72] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*. Cambridge University Press, 2012.

[73] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[74] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[75] Gong Chen and Marc Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64(1-3):81–101, 1994.

[76] Rodolphe Jenatton, Julien Mairal, Francis R Bach, and Guillaume R Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, 2010.

[77] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.

[78] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[79] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[80] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[81] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. In *Advances in neural information processing systems*, pages 905–912, 2009.

[82] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[83] Huahua Wang and Arindam Banerjee. Online alternating direction method (longer version). *arXiv preprint arXiv:1306.3721*, 2013.

[84] Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *Information Theory, IEEE Transactions on*, 57(7):4689–4708, 2011.

[85] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[86] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, pages 224–244, 2008.

[87] Joseph K Bradley, Aapo Kyrola, Danny Bickson, and Carlos Guestrin. Parallel coordinate descent for l1-regularized loss minimization. *arXiv preprint arXiv:1105.5379*, 2011.

[88] Peter Richtárik and Martin Takáv. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[89] Malaquias Peña and Huug van den Dool. Consolidation of multimodel forecasts by ridge regression: Application to pacific sea surface temperature. *Journal of Climate*, 21(24):6521–6538, 2008.

[90] Masoud Hessami, Philippe Gachon, Taha BMJ Ouarda, and André St-Hilaire. Automated regression-based statistical downscaling tool. *Environmental Modelling & Software*, 23(6):813–834, 2008.

[91] Larry M McMillin, Lawrence J Crone, and David S Crosby. Adjusting satellite radiances by regression with an orthogonal transformation to a prior estimate. *Journal of Applied Meteorology*, 28(9):969–975, 1989.

[92] Hasan Tatli, H Nüzhet Dalfes, and S Sibel Mentes. Surface air temperature variability over turkey and its connection to large-scale upper air circulation via multivariate techniques. *International Journal of Climatology*, 25(3):331–350, 2005.

[93] U. M. Fayyad, S. G. Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. In *Advances in Knowledge Discovery and Data Mining*. 1996.

[94] Lavrac et. al, editor. *Intelligent Data Analysis in Medicine and Pharmacology*. Kluwer, 1997.

[95] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Info. Th.*, 5112:4203–4215, 2005.

[96] L. Evers and C.M. Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, 2008.

[97] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proc. IEE*, 98(6):1031–1044, 2010.

[98] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, pages 487–494, June 2010.

[99] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1999.

[100] A. Tsonis, G. Wang, K. Swanson, F. Rodrigues, and L. Costa. Community structure and dynamics in climate networks. *Climate Dynamics*, 2010.

[101] F. Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 9:1179–1225, 2008.

[102] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2002.

[103] S. Chatterjee. An error bound in the Sudakov-Fernique inequality. Technical report, Arxiv, 2008. http://arxiv.org/abs/math/0510424.

[104] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[105] Nikhil S Rao, Ben Recht, and Robert D Nowak. Universal measurement bounds for structured sparse signal recovery. In *AISTATS*, pages 942–950, 2012.

[106] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. In *NIPS*, pages 1466–1474, 2012.

[107] A. M. McDonald, M. Pontil, and D. Stamos. New Perspectives on k-Support and Cluster Norms. *ArXiv e-prints*, March 2014, 1403.1481.

[108] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[109] Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 1987.

[110] John P Buonaccorsi. *Measurement error: models, methods, and applications*. CRC Press, 2010.

[111] Mathieu Rosenbaum and Alexandre B Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[112] Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[113] Mathieu Rosenbaum and Alexandre B. Tsybakov. Improved Matrix Uncertainty Selector. *arXiv:1112.4413*, 2011.

[114] Yudong Chen and Constantine Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *Proceedings of The 30th International Conference on Machine Learning*, pages 383–391, 2013.

[115] Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B. Tsybakov. An $\{l_1, l_2, l_\infty\}$-Regularization Approach to High-Dimensional Errors-in-variables Models. *arXiv:1412.7216*, 2014.

[116] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[117] Eunho Yang, Aurélie C Lozano, and Pradeep K Ravikumar. Elementary estimators for graphical models. In *Advances in Neural Information Processing Systems*, pages 2159–2167, 2014.

[118] Eunho Yang, Aurelie Lozano, and Pradeep Ravikumar. Elementary estimators for high-dimensional linear regression. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 388–396, 2014.

[119] D. P. Chambers, B. D. Tapley, and R. H. Stewart. Anomalous warming in the indian ocean coincident with el niño. *J. of Geoph. Res.*, 104:3035–3047, 1999.

[120] Richard H Moss, Jae A Edmonds, Kathy A Hibbard, Martin R Manning, Steven K Rose, Detlef P Van Vuuren, Timothy R Carter, Seita Emori, Mikiko Kainuma, Tom Kram, et al. The next generation of scenarios for climate change research and assessment. *Nature*, 463(7282):747–756, 2010.

[121] David A Randall, Richard A Wood, Sandrine Bony, Robert Colman, Thierry Fichefet, John Fyfe, Vladimir Kattsov, Andrew Pitman, Jagadish Shukla, Jayaraman Srinivasan, et al. Climate models and their evaluation. In *Climate Change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR)*, pages 589–662. Cambridge University Press, 2007.

[122] I-S Kang, K Jin, B Wang, K-M Lau, J Shukla, V Krishnamurthy, S Schubert, D Wailser, W Stern, A Kitoh, et al. Intercomparison of the climatological variations of asian summer monsoon precipitation simulated by 10 gcms. *Climate Dynamics*, 19(5-6):383–395, 2002.

[123] Xuejie Gao, Ying Shi, Ruiyan Song, Filippo Giorgi, Yongguang Wang, and Dongfeng Zhang. Reduction of future monsoon precipitation over china: Comparison between a high resolution rcm simulation and the driving gcm. *Meteorology and Atmospheric Physics*, 100(1-4):73–86, 2008.

[124] C Piani, JO Haerter, and E Coppola. Statistical bias correction for daily precipitation in regional climate models over europe. *Theoretical and Applied Climatology*, 99(1-2):187–192, 2010.

[125] K. Steinhaeuser, N. Chawla, and A. Ganguly. Comparing predictive power in climate data: Clustering matters. In *Advances in Spatial and Temporal Databases*, volume 6849, pages 39–55, 2011.

[126] Nils Chr Stenseth, Geir Ottersen, James W Hurrell, Atle Mysterud, Mauricio Lima, Kung-Sik Chan, Nigel G Yoccoz, and Bjorn Adlandsvik. Studying climate effects on ecology through the use of climate indices: the north atlantic oscillation, el nino southern oscillation and beyond. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1529):2087–2096, 2003.

[127] Saurabh V Pendse, Isaac K Tetteh, Fredrick HM Semazzi, Vipin Kumar, and Nagiza F Samatova. Toward data-driven, semi-automatic inference of phenomenological physical models: Application to eastern sahel rainfall. In *SDM*, pages 35–46, 2012.

[128] Rob Allan, Janette Lindesay, David Parker, et al. El nino: Southern oscillation and climatic variability. 1996.

[129] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[130] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

[131] Bryan FJ Manly. *Randomization, bootstrap and Monte Carlo methods in biology*, volume 70. CRC Press, 2006.

[132] Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *The Journal of Machine Learning Research*, 11:1833–1863, 2010.

[133] MJ Menne, CN Williams Jr, and RS Vose. United states historical climatology network (ushcn) version 2 serial monthly dataset. *Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee*, 2010.

[134] Fedor Mesinger, Geoff DiMego, Eugenia Kalnay, Kenneth Mitchell, Perry C Shafran, Wesley Ebisuzaki, Dusan Jovic, Jack Woollen, Eric Rogers, Ernesto H Berbery, et al. North american regional reanalysis. *Bulletin of the American Meteorological Society*, 87(3):343–360, 2006.

[135] A. Ganguly, E. Kodra, S. Chatterjee, A. Banerjee, and H. Najm. Computational data sciences for actionable insights on climate extremes and uncertainty. *Computational Intelligent Data Analysis for Sustainable Development*, page 1127, 2013.

[136] S. C. Liu, C. Fu, C. Shiu, J. Chen, and F. Wu. Temperature dependence of global precipitation extremes. *Geophysical Research Letters*, 36(17):L17702, 2009.

[137] Thomas A Niziol, Warren R Snyder, and Jeff S Waldstreicher. Winter weather forecasting throughout the eastern united states. part iv: Lake effect snow. *Weather and Forecasting*, 10(1):61–77, 1995.

[138] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 2013.

Appendix A

Proof of Theorems in Chapter 4

A.1 Proof of Theorem 3

Statement of Theorem: *Suppose the design matrix* $\mathbf{X}$ *consists of i.i.d. Gaussian entries with zero mean variance 1, and we solve the optimization problem* (4.1) *with*

$$\lambda_p \geq c\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] \ . \tag{A.1}$$

Then, with probability at least $(1 - \eta_1 \exp(-\eta_2 n))$*, we have*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{4c\Psi_{\mathcal{R}}\omega(\Omega_{\mathcal{R}})}{\kappa_{\mathcal{L}}\sqrt{n}} \ , \tag{A.2}$$

where $\omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})$ *is the Gaussian width of the intersection of* $\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)$ *and the unit spherical shell* $\mathbb{S}^{p-1}$*,* $\omega(\Omega_{\mathcal{R}})$ *is the Gaussian width of the unit norm ball,* $\kappa_{\mathcal{L}} > 0$ *is the gain given by*

$$\kappa_{\mathcal{L}} = \frac{1}{n}\left(\ell_n - \omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})\right)^2 \ , \tag{A.3}$$

$\Psi_{\mathcal{R}} = \sup_{\Delta \in \mathcal{T}_{\mathcal{R}}} \mathcal{R}(\Delta)/\|\Delta\|_2$ *is a norm compatibility factor,* ℓ_n *is the expected length of a length* n *i.i.d. standard Gaussian vector with* $\frac{n}{\sqrt{n+1}} < \ell_n < \sqrt{n}$*, and* $c > 1, \eta_1, \eta_2 > 0$ *are constants.*

Proof: We use the following lemma for the proof.

Lemma 6 *Suppose we solve the minimization problem* (4.1) *with* $\lambda_p \geq \mathcal{R}^*\left(\mathbf{X}^T\mathbf{w}\right)$*. Then the error vector* $\hat{\Delta}$ *belongs to the set*

$$\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*) := \text{cone}\left\{\Delta \in \mathbb{R}^p \ : \ \mathcal{R}(\boldsymbol{\theta}^* + \Delta) \leq \mathcal{R}(\boldsymbol{\theta}^*)\right\} \ , \tag{A.4}$$

and the error $\hat{\Delta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ satisfies the following bound

$$\mathcal{R}^* \left(\mathbf{X}^T \mathbf{X} \hat{\Delta} \right) \leq 2\lambda_p \tag{A.5}$$

Proof: By our choice of λ_p, both $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ lie in the feasible set of (4.1) , and by optimality of $\hat{\boldsymbol{\theta}}$,

$$\mathcal{R} \left(\boldsymbol{\theta}^* + \hat{\Delta} \right) = \mathcal{R}(\hat{\boldsymbol{\theta}}) \leq \mathcal{R}(\boldsymbol{\theta}^*) \, . \tag{A.6}$$

Also, by triangle inequality

$$\mathcal{R}^* \left(\mathbf{X}^T \mathbf{X} \hat{\Delta} \right) = \mathcal{R}^* \left(\mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \right) \tag{A.7}$$

$$\leq \mathcal{R}^* \left(\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}^*) \right) + \mathcal{R}^* \left(\mathbf{X}^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) \right) \leq 2\lambda_p \, . \tag{A.8}$$

■

Now, note that $\mathbf{X}$ and $\mathbf{w}$ are independent and we can rewrite

$$\mathbf{E}_{\mathbf{X},\mathbf{w}} \left[\mathcal{R}^*(\mathbf{X}^T \mathbf{w}) \right] = \mathbf{E}_{\mathbf{w}} \left[\mathbf{E}_{\mathbf{X}} \left[\mathcal{R}^*(\mathbf{X}^T \mathbf{w}) | \mathbf{w} \right] \right] = \mathbf{E}_{\mathbf{w}} \left[\|\mathbf{w}\|_2 \mathbf{E}_{\mathbf{X}} \left[\mathcal{R}^* \left(\mathbf{X}^T \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \right) | \mathbf{w} \right] \right] . \tag{A.9}$$

Since $\mathbf{w}/\|\mathbf{w}\|_2$ is an isotropic unit vector uniformly distributed over the surface of the unit sphere, $\left(\mathbf{X}^T \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \right) = \mathbf{g}$ is an i.i.d. $\mathcal{N}(0,1)$ Gaussian vector. Therefore

$$\mathbf{E}_{\mathbf{X},\mathbf{w}} \left[\mathcal{R}^*(\mathbf{X}^T \mathbf{w}) \right] = \mathbf{E}_{\mathbf{w}} [\|\mathbf{w}\|_2] \mathbf{E}_{\mathbf{g}} [\mathcal{R}^*(\mathbf{g})] \, . \tag{A.10}$$

Also, note that $\mathcal{R}^*(\cdot)$ is Lipschitz continuous with Lipschitz constant of 1 w.r.t. the norm $\mathcal{R}^*$, and hence by Gaussian concentration of Lipschitz functions [138],

$$\mathbb{P} \left(\mathcal{R}^*(\mathbf{g}) \geq \mathbf{E}_{\mathbf{g}} [\mathcal{R}^*(\mathbf{g})] + \tau \right) \leq \exp \left[-\frac{\tau^2}{2} \right] \, , \tag{A.11}$$

and similarly $\|\mathbf{w}\|_2 \leq \ell_n + \tau$ with probability at least $1 - \exp(-\tau^2/2)$, where $\frac{n}{\sqrt{n+1}} \leq \ell_n \leq \sqrt{n}$ is the expected length of $\mathbf{w}$. Therefore, for some $c > 1$ choosing $\lambda_p \geq c\mathbf{E} \left[\mathcal{R}^*(\mathbf{X}^T \mathbf{w}) \right] = c \, \ell_n \mathbf{E}_{\mathbf{g}} [\mathcal{R}^*(\mathbf{g})]$ implies that

$$\mathbb{P} \left(\lambda_p \geq \mathcal{R}^*(\mathbf{X}^T \mathbf{w}) \right) \geq \left(1 - \exp \left[-\frac{c_1 \mathbf{E}_{\mathbf{g}}^2 [\mathcal{R}^*(\mathbf{g})]}{2} \right] \right) \left(1 - \exp \left[-\frac{c_2 \ell_n^2}{2} \right] \right) = 1 - \eta_1' \exp(-\eta_2' n) \, , \tag{A.12}$$

for some constant $c_1, c_2, \eta_1', \eta_2' > 0$. Further, note that $\mathbf{E}_{\mathbf{g}}[\mathcal{R}^*(\mathbf{g})] = \omega(\Omega_{\mathcal{R}})$, the Gaussian width of the unit ball of norm $\mathcal{R}$.

Also, from Lemma 6, we have

$$\mathcal{R}^*\left(\mathbf{X}^T\mathbf{X}\hat{\Delta}\right) \leq 2\lambda_p \tag{A.13}$$

Now, note that

$$\|\mathbf{X}\hat{\Delta}\|_2^2 = \langle\hat{\Delta}, \mathbf{X}^T\mathbf{X}\hat{\Delta}\rangle \leq |\langle\hat{\Delta}, \mathbf{X}^T\mathbf{X}\hat{\Delta}\rangle| \leq \mathcal{R}(\hat{\Delta})\mathcal{R}^*\left(\mathbf{X}^T\mathbf{X}\hat{\Delta}\right) \leq 2\lambda_p\mathcal{R}(\hat{\Delta})\,, \tag{A.14}$$

where we have used Hölder's inequality, and the bound $\mathcal{R}^*\left(\mathbf{X}^T\mathbf{X}\hat{\Delta}\right) \leq 2\lambda_p$ from above.

Next, we use Gordon's theorem, which states that for $\mathbf{X}$ with i.i.d. Gaussian $(0,1)$ entries,

$$\mathbf{E}\left[\min_{\mathbf{z}\in\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1}} \|\mathbf{X}\mathbf{z}\|_2\right] \geq \ell_n - \omega\left(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1}\right)\,, \tag{A.15}$$

where ℓ_n is the expected length of an i.i.d. Gaussian random vector of length n, and $\omega\left(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1}\right)$ is the Gaussian width of the set $\Omega = \left(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1}\right)$. Now, since the function $\mathbf{X} \to \min_{\mathbf{z}\in\Omega}\|\mathbf{X}\mathbf{z}\|_2$ is Lipschitz continuous with constant 1 over the set Ω, we can use Gaussian concentration of Lipschitz functions [138] to obtain

$$\|\mathbf{X}\Delta\|_2 \geq \frac{1}{2}\left(\ell_n - \omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1})\right)\|\Delta\|_2 \tag{A.16}$$

$$\Rightarrow \quad \frac{1}{\sqrt{n}}\|\mathbf{X}\Delta\|_2 \geq \frac{\left(\ell_n - \omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1})\right)}{2\sqrt{n}}\|\Delta\|_2 \tag{A.17}$$

$$\Rightarrow \quad \frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq \frac{\kappa_{\mathcal{L}}}{2}\|\Delta\|_2^2\,, \tag{A.18}$$

with probability greater than $1 - \exp\left(-\frac{1}{8}\left(\ell_n - \omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1})\right)^2\right) = 1 - \eta_1''\exp(-\eta_2''n)$, where $\kappa_{\mathcal{L}} = \left(\ell_n - \omega(\mathcal{T}_{\mathcal{R}}(\boldsymbol{\theta}^*)\cap\mathbb{S}^{p-1})\right)^2/n > 0$ is the *gain*, and $\eta_1'', \eta_2'' > 0$ are constants .

Combining (A.18) and (A.14), and using the choice of λ_p, we obtain

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 = \|\hat{\Delta}\|_2 \leq \frac{4c\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right]}{\kappa_{\mathcal{L}}n}\frac{\mathcal{R}(\Delta)}{\|\Delta\|_2} \leq \frac{4c\Psi_{\mathcal{R}}\omega(\Omega_{\mathcal{R}})}{\kappa_{\mathcal{L}}\sqrt{n}} \tag{A.19}$$

with probability greater than $(1-\eta_1'\exp(-\eta_2'n))(1-\eta_1''\exp(-\eta_2''n)) = 1 - \eta_1\exp(-\eta_2 n)$, for constants η_1, η_2 where

$$\Psi_{\mathcal{R}} = \sup_{\Delta\in\mathcal{T}_{\mathcal{R}}}\frac{\mathcal{R}(\Delta)}{\|\Delta\|_2}\,. \tag{A.20}$$

The statement of the theorem follows. ∎

A.2 Proof of Theorem 4

Statement of Theorem: *For the k-support norm Generalized Dantzig Selection problem* (4.19)*, we obtain*

$$\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w})\right] \leq \sqrt{n}\left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right) \tag{A.21}$$

$$\omega(\Omega_{\mathcal{R}}) \leq \left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right) \tag{A.22}$$

$$\omega(\mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})^2 \leq \left(\sqrt{2k\log\left(p - k - \left\lceil\frac{s}{k}\right\rceil + 2\right)} + \sqrt{k}\right)^2 \cdot \left\lceil\frac{s}{k}\right\rceil + s\,. \tag{A.23}$$

Proof: We first illustrate that the k-support norm is an atomic norm, and then prove Theorem 4.

A.2.1 k-Support norm as an Atomic Norm

Here we show that k-support norm satisfies the definition of atomic norms [69]. Consider $\mathcal{G}_j$ to be the set of all subsets of $\{1, 2, \ldots, p\}$ of size j, so that

$$\mathcal{G}^{(k)} = \{\mathcal{G}_j\}_{j=1}^{k}\ . \tag{A.24}$$

For every j, consider the set

$$\mathcal{A}_j = \{\mathbf{w}\ :\ \|(\mathbf{w}_{G_j})\|_2 = 1,\ G_j \in \mathcal{G}_j,\ \mathbf{w}_i = \frac{1}{\sqrt{j}},\ \forall i \in G_j,\ \mathbf{w}_i = 0, \forall i \notin G_j\}\,, \tag{A.25}$$

corresponding to $\mathcal{G}_j$, and the union of such sets

$$\mathcal{A} = \{\mathcal{A}_j\}_{j\in\{1,\ldots,k\}}\ . \tag{A.26}$$

Note that since every non-zero element in a vector in $\mathcal{A}_j$ is $\frac{1}{\sqrt{j}}$, such an element cannot be represented as a convex combination of elements of the set $\mathcal{A}_l,\ l < j$, whose non-zero elements are $\frac{1}{\sqrt{l}}$. Therefore none of the elements $\mathbf{w}$ in the set $\mathcal{A}$ lies in the convex hull of the other elements $\mathcal{A} \setminus \{\mathbf{w}\}$. Further, note that

$$\text{conv}(\mathcal{A}) = C_k\ , \tag{A.27}$$

and the k-support norm defines the gauge function of the $\mathcal{A}$. Thus the k-support norm is an atomic norm.

A.2.2 The Error set and its Gaussian width

Note that the cardinality of the set $\mathcal{G}^{(k)}$ is

$$M = \binom{p}{k} + \binom{p}{k-1} + \binom{p}{k-2} + \cdots + \binom{p}{1} \tag{A.28}$$

The error set is given by

$$\mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*) = \text{cone}\{\Delta \in \mathbb{R}^p \ : \ \|\Delta + \boldsymbol{\theta}^*\|_k^{sp} \leq \|\boldsymbol{\theta}^*\|_k^{sp}\} \, . \tag{A.29}$$

Note that this set is a cone, and we can define the *normal* cone of this set as

$$\mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*) = \{\mathbf{u} \ : \ \langle \mathbf{u}, \Delta \rangle \leq 0, \ \forall \Delta \in \mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*)\} \tag{A.30}$$

$$\tag{A.31}$$

The following proposition, shown in [105], shows that the normal cone can be written in terms of the dual norm of the k-support norm.

Proposition 1 *The normal cone to the tangent cone defined in* (A.29) *can written as*

$$\mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*) = \{\mathbf{u} \ : \ \exists t > 0 \ \textit{s.t.} \ \langle \mathbf{u}, \boldsymbol{\theta}^* \rangle = t \|\boldsymbol{\theta}^*\|_k^{sp}, \ \|\mathbf{u}\|_k^{sp^*} \leq t\} \, . \tag{A.32}$$

We provide a simple proof of this statement for our case for ease of understanding.

Proof: We re-write the definition of the normal cone in terms of the estimated parameter $\hat{\boldsymbol{\theta}}$ as

$$\mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*) = \{\mathbf{u} \in \mathbb{R}^p \ : \ \langle \mathbf{u}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq 0, \forall \boldsymbol{\theta} - \boldsymbol{\theta}^* \in \mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*)\} \, . \tag{A.33}$$

Note that this means that $\mathbf{u} \in \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*)$ if and only if

$$\langle \mathbf{u}, \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle \leq 0, \ \ \forall \|\boldsymbol{\theta}\|_k^{sp} \leq \|\boldsymbol{\theta}^*\|_k^{sp} \tag{A.34}$$

$$\Rightarrow \langle \mathbf{u}, \boldsymbol{\theta} \rangle \leq \langle \mathbf{u}, \boldsymbol{\theta}^* \rangle \ \ \forall \|\boldsymbol{\theta}\|_k^{sp} \leq \|\boldsymbol{\theta}^*\|_k^{sp} \, . \tag{A.35}$$

Now, we claim that $\langle \mathbf{u}, \boldsymbol{\theta}^* \rangle \geq 0$ for all such $\mathbf{u}$. This can be shown as follows. Assume the contrary, i.e. there exists a $\hat{\mathbf{u}} \in \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*)$ such that $\langle \hat{\mathbf{u}}, \boldsymbol{\theta}^* \rangle < 0$. Now, noting that $(-\boldsymbol{\theta}^*) \in \mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*)$, we have

$$\langle \hat{\mathbf{u}}, -\boldsymbol{\theta}^* \rangle = -\langle \hat{\mathbf{u}}, \boldsymbol{\theta}^* \rangle > 0 \, , \tag{A.36}$$

so that $\hat{\mathbf{u}} \notin \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*)$, which is a contradiction, and the claim follows.

Therefore, we can write

$$\langle \mathbf{u}, \boldsymbol{\theta}^* \rangle = t \|\boldsymbol{\theta}^*\|_k^{sp} \tag{A.37}$$

for some $t \geq 0$. Then, $\mathbf{u} \in \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*)$ if and only if

$$\exists t \geq 0 \ , \ \langle \mathbf{u}, \boldsymbol{\theta}^* \rangle = t \|\boldsymbol{\theta}^*\|_k^{sp} \ \ , \ \ \langle \mathbf{u}, \boldsymbol{\theta} \rangle \leq t \|\boldsymbol{\theta}^*\|_k^{sp} \ \ \forall \|\boldsymbol{\theta}\|_k^{sp} \leq \|\boldsymbol{\theta}^*\|_k^{sp} \ . \tag{A.38}$$

Since

$$\langle \mathbf{u}, \boldsymbol{\theta} \rangle \leq t \|\boldsymbol{\theta}^*\|_k^{sp}, \ \ \forall \|\boldsymbol{\theta}\|_k^{sp} \leq \|\boldsymbol{\theta}^*\|_k^{sp} \ \Rightarrow \ \|\mathbf{u}\|_k^{sp^*} \leq t \ , \tag{A.39}$$

the statement follows. ∎

The k-support norm can be thought of as a group sparse norm with overlaps, such as been dealt with in [105]. Therefore, we can utilize some of the analysis techniques developed in [105], specialized to the structure of the k-support norm. We begin by stating a theorem which enables us to bound the Gaussian width of the error set. Henceforth, we write $\mathcal{N}_{\mathcal{A}} = \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*)$ and $\mathcal{T}_{\mathcal{A}} = \mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*)$ where the dependence on $\boldsymbol{\theta}^*$ is understood.

First, we define sets that involve the support set of $\boldsymbol{\theta}^*$. Let us define the set $\mathcal{G}^* \subseteq \mathcal{G}^{(k)}$ to be the set of all groups in $\mathcal{G}^{(k)}$ which overlap with the support of $\boldsymbol{\theta}^*$, i.e.

$$\mathcal{G}^* = \{G \in \mathcal{G}^{(k)} \ : \ G \cap \text{supp}(\boldsymbol{\theta}^*) \neq \emptyset\} \ . \tag{A.40}$$

Let S be the union of all groups in $\mathcal{G}^*$, i.e. $S = \bigcup_{G \in \mathcal{G}^*} G$, and the size of S be $|S| = s$. We are going to use three lemmas in order to prove the above bound. The first lemma, proved in [69], upper bounds the Gaussian width by an expected distance to the normal cone as follows.

Lemma 7 ([69] Proposition 3.6) *Let $\mathbb{C}$ be any nonempty convex in $\mathbb{R}^p$, and $\mathbf{g} \sim \mathcal{N}(0, I_p)$ be a random gaussian vector. Then*

$$\omega(\mathbb{C} \cap \mathbb{S}^{p-1}) \leq \mathbf{E}_{\mathbf{g}}[\textit{dist}(\mathbf{g}, \mathbb{C}^*)] \ , \tag{A.41}$$

where $\mathbb{C}^*$ is the polar cone of $\mathbb{C}$.

Note that $\mathcal{N}_{\mathcal{A}}$ is the polar cone of $\mathcal{T}_{\mathcal{A}}$ by definition. Therefore, using Jensen's inequality, we obtain

$$\omega(\mathcal{T}_{\mathcal{A}} \cap \mathbb{S}^{p-1})^2 \leq \mathbf{E}_{\mathbf{g}}^2[\text{dist}(\mathbf{g}, \mathcal{N}_{\mathcal{A}})] \leq \mathbf{E}_{\mathbf{g}}[\text{dist}(\mathbf{g}, \mathcal{N}_{\mathcal{A}})^2] \leq \mathbf{E}_{\mathbf{g}}[\|\mathbf{g} - \mathbf{z}(\mathbf{g})\|_2^2] \ , \tag{A.42}$$

where $\mathbf{z}(\mathbf{g}) \in \mathcal{N}_{\mathcal{A}}$ is a (random) vector constructed to lie always in the normal cone. The construction proceeds as follows.

Constructing $\mathbf{z}(\mathbf{g})$: Note that $\boldsymbol{\theta}^*_{S^c} = 0$. Let us choose a vector $\mathbf{v} \in \mathcal{N}_{\mathcal{A}}$ such that

$$\|\mathbf{v}\|_k^{sp^*} = 1 \text{ and } \mathbf{v}_{S^c} = 0 \, . \tag{A.43}$$

We can choose an appropriately scaled $\mathbf{v}$ so that

$$\langle \mathbf{v}, \boldsymbol{\theta}^* \rangle = \|\boldsymbol{\theta}^*\|_k^{sp} \, , \tag{A.44}$$

and let us write without loss of generality $\mathbf{v} = [\mathbf{v}_S \;\; \mathbf{v}_{S^c}]$.

Next, let $\mathbf{g} \sim \mathcal{N}(0, I_p)$, and write $\mathbf{g} = [\mathbf{g}_S \;\; \mathbf{g}_{S^c}]$. We define the quantity

$$t(\mathbf{g}) = \max\Big\{ \|\mathbf{g}_G\|_2 \, : \, G \in \mathcal{G}^{(k)}, G \subseteq S^c \Big\} = \max\left\{ \left(\sum_{i \in G} \mathbf{g}_i^2 \right)^{\frac{1}{2}} \, : \, G \in \mathcal{G}^{(k)}, G \subseteq S^c \right\} \, , \tag{A.45}$$

and let $\mathbf{z} = \mathbf{z}(\mathbf{g}) = [\mathbf{z}_S \;\; \mathbf{z}_{S^c}]$ such that

$$\mathbf{z}_S = t(\mathbf{g}) \mathbf{v}_S, \quad \mathbf{z}_{S^c} = \mathbf{g}_{S^c} \, . \tag{A.46}$$

Note that

$$\langle \mathbf{z}, \boldsymbol{\theta}^* \rangle = t(\mathbf{g}) \langle \mathbf{v}_S, \boldsymbol{\theta}^*_S \rangle = t(\mathbf{g}) \|\boldsymbol{\theta}^*\|_k^{sp} \, , \tag{A.47}$$

and

$$\|\mathbf{z}\|_k^{sp^*} = \max\Big\{ \|\mathbf{z}_G\|_2 \, : \, G \in \mathcal{G}^{(k)} \Big\} \tag{A.48}$$

$$= \max\Big\{ \max\{\|\mathbf{z}_G\|_2 \, : \, G \in \mathcal{G}^{(k)}, G \subseteq S\} \, , \; \max\{\|\mathbf{z}_G\|_2 \, : \, G \in \mathcal{G}^{(k)}, G \subseteq S^c\} \Big\} \tag{A.49}$$

$$\overset{(a)}{=} \max\Big\{ t(\mathbf{g}) \|\mathbf{v}\|_k^{sp^*} \, , \; t(\mathbf{g}) \Big\} \tag{A.50}$$

$$= t(\mathbf{g}) \tag{A.51}$$

where (a) follows from the definition of $t(\mathbf{g})$ and the fact that

$$\max\{\|\mathbf{z}_G\|_2 \, : \, G \in \mathcal{G}^{(k)}, G \subseteq S\} = t(\mathbf{g}) \max\{\|\mathbf{v}_G\|_2 \, : \, G \in \mathcal{G}^{(k)}, G \subseteq S\} = t(\mathbf{g}) \|\mathbf{v}\|_k^{sp^*} \, , \tag{A.52}$$

and since $\|\mathbf{v}\|_k^{sp^*} = 1$. Therefore, $\mathbf{z}(\mathbf{g}) \in \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*)$ by definition in (A.32) .

In order to upper bound the expectation of $t(\mathbf{g})$, we use the following comparison inequality from [105].

Lemma 8 ([105] Lemma 3.2) *Let $q_1, q_2, \ldots, q_L$ be L, χ-squared random variables with d degrees of freedom. Then*

$$\mathbf{E}\left[\max_{1\leq i\leq L} q_i\right] \leq \left(\sqrt{2\log L} + \sqrt{d}\right)^2 . \tag{A.53}$$

Last, we prove an upper bound on the expected value of $t(\mathbf{g})$, as shown in the following lemma.

Lemma 9 *Consider $\mathcal{G}^* \subseteq \mathcal{G}^{(k)}$ to be the set of groups intersecting with the support of $\boldsymbol{\theta}^*$, and let S be the union of groups in $\mathcal{G}^*$, such that $s = |S|$. Then,*

$$\mathbf{E}_{\mathbf{g}}[t(\mathbf{g})^2] \leq \left(\sqrt{2k\log\left(p-k-\left\lceil\frac{s}{k}\right\rceil+2\right)} + \sqrt{k}\right)^2 . \tag{A.54}$$

Proof: Note that

$$\mathbf{E}_{\mathbf{g}}[t(\mathbf{g})^2] = \mathbf{E}_{\mathbf{g}}\left[\left(\max\left\{\|\mathbf{g}_G\|_2 \ : \ G\in\mathcal{G}^{(k)}, G\subseteq S^c\right\}\right)^2\right] \tag{A.55}$$

$$\leq \mathbf{E}_{\mathbf{g}}\left[\max\left\{\|\mathbf{g}_G\|_2^2 \ : \ G\in\mathcal{G}^{(k)}, G\subseteq S^c\right\}\right] \tag{A.56}$$

Each term $\|\mathbf{g}_G\|_2^2$ is a χ-squared variable with at most k degrees of freedom. Since the set S has size s, the set $\mathcal{G}^*$ has to contain at least $s_k = \lceil\frac{s}{k}\rceil$ groups of size k. Therefore,

$$s = |S| \geq k + (s_k - 1) , \tag{A.57}$$

and therefore the size of its complement is upper bounded by

$$|S^c| \leq p - k - s_k + 1 . \tag{A.58}$$

Therefore the following inequality provides an upper bound on the number of groups involved in computing the maximum in (A.56)

$$\left|\left\{G\in\mathcal{G}^{(k)}, G\subseteq S^c\right\}\right| \leq \binom{p-k-s_k+1}{k} + \binom{p-k-s_k+1}{k-1} + \cdots + \binom{p-k-s_k+1}{1} \tag{A.59}$$

$$\leq (p-k-s_k+2)^k \tag{A.60}$$

where we have used the following inequality

$$\binom{n}{h} \leq \frac{n^h}{h!}, \ \ \forall n\geq h\geq 0 , \tag{A.61}$$

which also provides

$$\sum_{h=1}^{k} \binom{n}{h} \leq (n+1)^k \,. \tag{A.62}$$

Therefore, we can upper bound (A.56) using Lemma 8 as

$$\mathbf{E}_{\mathbf{g}}[t(\mathbf{g})^2] \leq \mathbf{E}_{\mathbf{g}}\Big[\max\Big\{ \|\mathbf{g}_G\|_2^2 \,:\, G \in \mathcal{G}^{(k)}, G \subseteq S^c \Big\} \Big] \tag{A.63}$$

$$\leq \left(\sqrt{2 \log\left((p-k-\left\lceil \frac{s}{k} \right\rceil + 2)^k \right)} + \sqrt{k} \right)^2 \tag{A.64}$$

and the statement follows. ∎

Now we are ready to prove the upper bound on the Gaussian width. First, note that

$$\omega(\mathcal{T}_{\mathcal{A}}(\boldsymbol{\theta}^*) \cap \mathbb{S}^{p-1})^2 \leq \mathbf{E}_{\mathbf{g}}[\mathrm{dist}(\mathbf{g}, \mathcal{N}_{\mathcal{A}}(\boldsymbol{\theta}^*))^2] \tag{A.65}$$

$$\overset{(a)}{\leq} \mathbf{E}_{\mathbf{g}}[\|\mathbf{g} - \mathbf{z}(\mathbf{g})\|_2^2] \tag{A.66}$$

$$= \mathbf{E}_{\mathbf{w}}[\|\mathbf{z}_S - \mathbf{g}_S\|_2^2] \tag{A.67}$$

$$\overset{(b)}{=} \mathbf{E}[\|\mathbf{z}_S\|_2^2] + \mathbf{E}[\|\mathbf{g}_S\|_2^2] \tag{A.68}$$

$$\overset{(c)}{=} \mathbf{E}[t(\mathbf{g})^2] \cdot \|\mathbf{v}_S\|_2^2 + |S| \tag{A.69}$$

$$\overset{(d)}{\leq} \left(\sqrt{2k \log\left((p-k-\left\lceil \frac{s}{k} \right\rceil + 2) \right)} + \sqrt{k} \right)^2 \cdot \left\lceil \frac{s}{k} \right\rceil + s \,, \tag{A.70}$$

where (a) follows from the definition of distance to a set, (b) follows from the independence of $\mathbf{g}_S$ and $\mathbf{g}_{S^c}$, (c) follows from the fact that the expected length of an $|S|$ length random i.i.d. Gaussian vector is $\sqrt{|S|}$, and (d) follows since $|S| = \frac{ks}{k}$, and that $\|\mathbf{v}_S\|_2 \leq \sqrt{\left\lceil \frac{s}{k} \right\rceil} \|\mathbf{v}_S\|_k^{sp^*} = \sqrt{\left\lceil \frac{s}{k} \right\rceil}$. Thus inequality (A.23) follows. ∎

Next, we prove inequality (A.21). Let us denote $\mathbf{t} = \mathbf{X}^T \left(\frac{\mathbf{w}}{\|\mathbf{w}\|_2} \right)$, and note that $\mathbf{t} \sim \mathcal{N}(0, I_p)$. Also note that $\mathbf{E}\left[\mathcal{R}^*(\mathbf{X}^T\mathbf{w}) \right] = E[\|\mathbf{w}\|_2\|]\mathbf{E}[\mathcal{R}^*(\mathbf{t})]$, and

$$\|\mathbf{t}\|_k^{sp^*} = \max\{ \|\mathbf{t}_G\|_2 \,:\, G \in \mathcal{G}^{(k)} \} \,. \tag{A.71}$$

Therefore, we can use Lemma 8 in order to bound the expectation $\mathbf{E}[\|\mathbf{t}\|_k^{sp^*}]$ as

$$\mathbf{E}[\|\mathbf{t}\|_k^{sp^*}] = \mathbf{E}[\max\{\|\mathbf{t}_G\|_2 \ : \ G \in \mathcal{G}^{(k)}\}] \tag{A.72}$$

$$= \mathbf{E}[\max\{\|\mathbf{t}_G\|_2 \ : \ G \in \mathcal{G}^{(k)}, \ |G| = k\} \tag{A.73}$$

$$\leq \left(\sqrt{2\log\binom{p}{k}} + \sqrt{k}\right) \tag{A.74}$$

$$\leq \left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right) \ , \tag{A.75}$$

where we have used the following inequality obtained using Stirling's approximation

$$\binom{p}{k} \leq \left(\frac{pe}{k}\right)^k \ . \tag{A.76}$$

Therefore, inequality (A.21) follows, and by our choice of λ_p, with high probability, $\boldsymbol{\theta}^*$ lies in the feasible set.

Last, note that

$$\omega(\Omega_{\mathcal{R}}) = \mathbf{E}[\|\mathbf{t}\|_k^{sp^*}] \leq \left(\sqrt{2k\log\left(\frac{pe}{k}\right)} + \sqrt{k}\right) \ , \tag{A.77}$$

as proved above. ■