Title

Some Inference Problems in High-Dimensional Linear Models

Author

Lopes, Miles Edward

Publication Date

2015

Some Inference Problems in High-Dimensional Linear Models

by

Miles Edward Lopes

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, Chair
Professor Lior Pachter
Professor Martin J. Wainwright

Spring 2015

Some Inference Problems in High-Dimensional Linear Models

Abstract

Some Inference Problems in High-Dimensional Linear Models

by

Miles Edward Lopes

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter J. Bickel, Chair

During the past two decades, technological advances have led to a proliferation of high-dimensional problems in data analysis. The characteristic feature of such problems is that they involve large numbers of unknown parameters and relatively few observations. As the study of high-dimensional statistical models has developed, linear models have taken on a special status for their widespread application and extensive theory. Even so, much of the theoretical research on high-dimensional linear models has been concentrated on the problems of prediction and estimation, and many inferential questions regarding hypothesis tests and confidence intervals remain open.

In this dissertation, we explore two sets of inferential questions arising in high-dimensional linear models. The first set deals with the residual bootstrap (RB) method and the distributional approximation of regression contrasts. The second set addresses the issue of unknown sparsity in the signal processing framework of compressed sensing. Although these topics involve distinct methods and applications, the dissertation is unified by an overall focus on the interplay between model structure and inference. Specifically, our work is motivated by an interest in using inferential methods to confirm the existence of model structure, and in developing new inferential methods that have minimal reliance on structural assumptions.

The residual bootstrap method is a general approach to approximating the sampling distribution of statistics derived from estimated regression coefficients. When the number of regression coefficients p is small compared to the number of observations n, classical results show that RB consistently approximates the laws of contrasts obtained from least-squares coefficients. However, when $p/n \asymp 1$, it is known that there exist contrasts for which RB fails — when applied to least-squares residuals. As a remedy, we propose an alternative method that is tailored to regression models involving near low-rank design matrices. In this situation, we prove that resampling the residuals of a ridge regression estimator can alleviate some of the problems that occur for least-squares residuals. Notably, our approach does not depend on sparsity in the true regression coefficients. Furthermore, the assumption of a near low-rank design is one that is satisfied in many applications and can be inspected directly in practice.

In the second portion of the dissertation, we turn our attention to the subject of compressed sensing, which deals with the recovery of sparse high-dimensional signals from a limited number of linear measurements. Although the theory of compressed sensing offers strong recovery guarantees, many of its basic results depend on prior knowledge of the signal's sparsity level — a parameter that is rarely known in practice. Towards a resolution of this issue, we introduce a generalized family of sparsity parameters that can be estimated in a way that is free of structural assumptions. We show that our estimator is ratio-consisent with a dimension-free rate of convergence, and also derive the estimator's limiting distribution. In turn, these results make it possible to set confidence intervals for the sparsity level and to test the hypothesis of sparsity in a precise sense.

To my family

Contents

List of Figures

List of Tables

List of Symbols

$a_n \lesssim b_n$ For two sequences of real numbers a_n and b_n, the relation $a_n \lesssim b_n$ means that there is a constant $c > 0$, and a number $n_0 \geq 1$, such that $a_n \leq cb_n$ for all $n \geq n_0$.

$a_n = \mathcal{O}(b_n)$ This relation has the same meaning as $a_n \lesssim b_n$.

$a_n = o(b_n)$ For two sequences of real numbers a_n and b_n, the relation $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$.

$a_n \asymp b_n$ This relation means that both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold.

$\longrightarrow_P$ Convergence in probability.

$U_n = \mathcal{O}_P(1)$ If U_n is a sequence of random variables with $n = 1, 2, \ldots$, then this relation means that for any $\varepsilon > 0$, there is finite constant $M > 0$ such that the bound $\mathbb{P}(|U_n| \geq M) \leq \varepsilon$ holds for all $n \geq 1$.

$U_n = \mathcal{O}_P(V_n)$ If U_n and V_n are sequences of random variables with $n = 1, 2, \ldots$, then this relation means that there is another sequence of random variables R_n with $R_n = \mathcal{O}_P(1)$ such that $U_n = R_n V_n$.

$U_n = o_P(1)$ This relation means that $U_n \longrightarrow_P 0$.

$\xrightarrow{w}$ Weak convergence.

$\mathcal{L}(U|V)$ The law of a random variable U conditioned on a random variable V.

$d_2(\mu, \nu)$ The Mallows (Kantorovich) ℓ_2 distance between distributions μ and ν.

$d_{\mathrm{LP}}(\mu, \nu)$ The Lévy-Prohorov distance between distributions μ and ν

$\|x\|_q$ The ℓ_q norm, equal to $(\sum_{i=1}^p |x_i|^q)^{1/q}$ for $x \in \mathbb{R}^p$ and $q > 0$. When $q \in (0,1)$ the quantity is not a norm but but still referred to as such, and similarly for $\|x\|_0 = \text{card}\{j \in \{1,\dots,p\} : x_j \neq 0\}$.

$|||A|||_q$ The Schatten-q norm, equal to $\|\sigma(A)\|_q$ where $\sigma(A)$ is a vector containing the singular values of A. The quantity is only a proper norm for $q \geq 1$, and when $q = 0$, $|||A|||_0 = \text{rank}(A)$.

$\lambda_i(A)$ The ith largest eigenvalue of a real symmetric matrix A.

$\text{tr}(A)$ The trace of a matrix A.

$\|A\|_F$ The Frobenius norm of a matrix A, equal to $\sqrt{\sum_{ij} A_{ij}^2}$.

$\|A\|_{\text{op}}$ The operator norm (maximum singular value) of a matrix A.

$\ell^\infty(\mathcal{I})$ The space of bounded (real or complex) functions on an interval $\mathcal{I} \subset \mathbb{R}$.

$\mathscr{C}(\mathcal{I})$ The space of continuous (real or complex) functions on an interval $\mathcal{I} \subset \mathbb{R}$.

Acknowledgments

First and foremost, I would like to acknowledge my advisor Peter Bickel for his generosity of time and encouragement. I am especially fortunate to have benefited from his panoramic perspective on statistics, and the unlimited degree of freedom he provided me in pursuing research. It is difficult to imagine a better example to follow, on either a scientific or personal level.

With regard to my dissertation and qualifying exam, I would like to thank the members of my committees, which in addition to Peter, include Lior Pachter, Martin Wainwright, and Sandrine Dudoit. In particular, I give special thanks to Martin for supervising my masters degree in computer science, giving outstanding lectures, and offering his help over the course of my graduate study.

During my time in graduate school, I was fortunate enough to have participated in several internships and research visits. I thank Philip Kegelmeyer for hosting my practicum at Sandia National Laboratories in 2012. His enthusiasm and intuition for ensemble methods has had a substantial influence on my research trajectory. Next, I thank Zaïd Harchaoui and Anatoli Juditsky for the great experience of collaborating with them at INRIA in Grenoble during the spring of 2013. I especially thank Zaïd for being an outstanding mentor and friend during the last several years. More recently, I had the opportunity to work as a summer intern at Google, and I thank Aiyou Chen for being an excellent host and advisor. I learned a great deal about what it means to solve large problems from Aiyou, as well as many other members of the QM team at Google.

My graduate study was influenced by numerous colleagues and professors — in fact, too many to name in full, and I apologize in advance to those who I have left out. I thank Jeff Regier and Miki Racz for being great officemates, as well Aditya Guntuboyina and Hongwei Li for always being available to chat. I owe Laurent Jacob a great deal of thanks for having the patience to help me get my research off the ground, and for being a wonderful collaborator on my first paper. I thank Ani Adhikari for being a terrific source of wisdom regarding teaching, and for her help at numerous points of graduate school. Also, I thank Ryan Lovett, La Shana Porlaris, Mary Melinn for making the department run so smoothly and always having time to lend a hand.

Outside of Berkeley, I was supported by many friends, teachers, and organizations. I thank Alan Jern and Deran Atamian for being true friends and sources of perspective. I also thank Richard Elman, Peter Petersen, and Michael Gutperle at UCLA for their encouragement, and for stimulating my interest in mathematical research at an early stage. With regard to financial support, I thank the DOE CSGF and NSF GRFP programs for providing me with the time and freedom to pursue research with so few constraints.

Most of all, I thank my family for their unconditional support and encouragement at every step of the way.

Chapter 1

Introduction

In this chapter, we first introduce the subject of high-dimensional statistics in Section 1, and then discuss the motivations of the dissertation in Section 2. Since the dissertation will address two distinct areas of high-dimensional statistics, namely bootstrap methods and compressed sensing, we provide some background on each of these topics in Sections 1.3 and 1.4. Lastly, in Section 1.5, we conclude by describing the specific contributions of the dissertation and outlining its remaining chapters.

1.1 High-dimensional statistics

The subject of high-dimensional statistics deals with problems where relatively few observations are available to estimate a large number of parameters. Although problems of this type have always been a part of data analysis, their prevalence in scientific applications is relatively new, and roughly speaking, the subject did not emerge as a major branch of statistical research until the last 15 years. This situation is not an historical accident. Only in the last couple of decades has computational power made it feasible manipulate high-dimensional datasets in real time. Similarly, the rise of computation has also enabled new data acquisition technologies that can measure large numbers of features simultaneously — making high-dimensional data ubiquitous in many areas of science.

Examples

Below, we briefly describe some examples of high-dimensional data. Our small set of examples reflects just a handful of applications that have received significant attention from statisticians in recent years. This list does not begin to scratch the surface of the numerous areas where high-dimensional data occur. Some references that give a more wide-ranging view of high-dimensional data include the National Research Council's 2013 report *Frontiers of Massive Data Analysis* [Cou13], as well as the National Science Foundation's report ensuing from their 2007 workshop on *Discovery in Complex or Massive Datasets: Common Statistical Themes* [Fou07]

- **Internet commerce.** Recommender systems are one of the most widely used technologies for selling products online [RV97]. In many cases, such systems can be thought of in terms of a large matrix, with rows indexed by users and columns indexed by products. Each entry of the matrix is a score measuring a user's preference for a product, and these entries are the parameters of interest. The observed data consist only of a relatively small number of preferences generated by users on products they have already purchased, and the matrix completion problem of estimating *all* of these preferences is clearly high-dimensional. During the last several years, the so-called Netflix Prize attracted considerable interest to this problem in the context of movie recommendations, and a large stream of research has ensued. We refer the reader to the papers [CR09] [F+12] and the references therein.

- **Networks.** In recent years, there has been an explosion of research on the statistical analysis of networks, especially with attention to information networks, biological networks, and social networks [Kol09]. To mention just one example concretely, consider the application of internet security, which often deals with the task of monitoring the traffic of packet exchanges across a network of internet protocol (IP) addresses. In particular, it is of basic interest to detect anomalous traffic patterns [LR09]. If there are p edges connecting the addresses, and we use a vector $\Delta(t) \in \mathbb{R}^p$ to record the number of packets exchanged along each edge at a time point t, then it is natural to formulate the problem as detecting change points involving the p-dimensional time series $\Delta(t)$. To see the extreme role of dimensionality in this problem, note that many information networks have a number of nodes N on the order of several thousand, and the number of edges is often of order $p \asymp N^2$. Moreover, due to the computational constraints of counting packet exchanges, we may be limited to a number of observations $\Delta(1), \dots, \Delta(n)$ with $n \ll p$. Changepoint detection in this setting is extremely difficult, and due to the rising importance of internet security, research in this direction is likely to continue for some time to come.

- **Genomics.** The human genome consists of roughly fifteen thousand genes, and the expression level of any particular gene can be viewed as a numerical feature of an individual organism. Due to the fact that biological studies are often limited to several dozen individuals, it is not uncommon to encounter situations where the number of observations n is drastically less than their dimension p. An example of a problem in statistical genomics where dimensionality plays an especially prominent role is detecting a difference between a treatment population and a control population, i.e. the *two-sample test.* When a large number of genes are used to make such a comparison, it is relatively easy for "chance variation" across the genes to "explain away" any systematic difference between the two groups. In fact, the two-sample test was one of the earliest problems where the effect of dimensionality on statistical power was quantified in a precise way — as shown in a seminal paper [BS96]. In the years since this work, a substantial line of research has studied different approaches for combatting the loss of statistical power that occurs in high-dimensional testing problems. A review of the interactions between genomics and high-dimensional data analysis may be found in the paper [Bic+09].

Asymptotics

Much of the classical asymptotic theory of statistics is formulated in the following way. A set of i.i.d. observations $X_1, \dots, X_n$ is generated from a distribution $\mathbb{P}_\theta$ on a sample space $\mathcal{X}$, where θ is an unknown parameter in a subset of p-dimensional Euclidean space, $\Theta \subset \mathbb{R}^p$. Using the observations, an estimator $\widehat{\theta}$ is computed, and we are interested in understanding how well $\widehat{\theta}$ approximates θ as n becomes large. Traditionally, the analysis of $\widehat{\theta}$ is done under the assumption that both Θ and $\mathcal{X}$ remain fixed as $n \to \infty$, and that the entire sequence $\{X_i\}_{i=1}^{\infty}$ is generated from $\mathbb{P}_\theta$.

A drawback of the classical framework is that it is restricted to describing situations where the dimension p is small with respect to n. This limitation creates difficulties in practical problems, because for a particular dataset, where say $n = 100$ and $p = 37$, it is not always clear if p should be viewed as negligible in comparison to n. In turn, we may doubt whether the results of standard asymptotic calculations are relevant. Furthermore, in problems where $n < p$, the use of asymptotics with p held fixed is often entirely inappropriate. In such cases, where p is at least of the same order of magnitude as n, a problem is typically labelled as *high-dimensional*.[1]

In order to describe situations where p and n are both large, various alternatives to classical asymptotics have been proposed. Some of the earliest examples of asymptotics considered from this high-dimensional viewpoint include Huber's program of robust regression in the early 1970's [Hub73], as well as the work of Kolmogorov on discriminant analysis in the late 1960's [Ser08, see preface], [Aiv+89]. Outside of statistics, the use of high-dimensional asymptotics occurred even earlier during the 1950's in connection with mathematical physics and Wigner's analysis of the spectra of large random matrices [Wig58].

The most standard way of formalizing the notion of p and n diverging simultaneously is to consider a sequence of "growing models" ordered by a "latent index", say $\xi \in \mathbb{N}$. Specifically, for each $\xi = 1, 2, \dots$, we are interested in an unknown parameter $\theta = \theta(\xi)$ lying in a space whose dimension also varies with ξ, e.g. $\Theta(\xi) \subset \mathbb{R}^{p(\xi)}$. Furthermore, for each ξ, we observe a number $n = n(\xi)$ of i.i.d. samples $X_1, \dots, X_{n(\xi)}$ from a distribution $\mathbb{P}_{\theta(\xi)}$, creating an array whose rows are indexed by ξ. (The sample space of the X_i may also vary with ξ.) Under this formalism, the goal is to understand how an estimator $\widehat{\theta}(\xi)$ derived from $X_1, \dots, X_{n(\xi)}$ approximates $\theta(\xi)$ as $(p(\xi), n(\xi)) \to \infty$ with $\xi \to \infty$. For ease of notation, it has become

[1] Although it could be argued that *non-parametric* or *semi-parametric* problems are high-dimensional since θ may lie in a space of infinite dimension, the term high-dimensional is most often used in the context of *parametric* models involving a large but finite number of parameters.

standard in the statistics literature to suppress the latent index, and we will mostly follow this convention going forward.

As soon as we allow p and n tend to infinity together, it is often necessary choose their relative rates of growth in order to carry out calculations. While this choice might appear to be a technical detail, it is actually a fundamental demarcation point for different types of problems in high-dimensional statistics. Typically, this choice is specified in terms of the ratio p/n. The limiting value of the ratio (or its rate of growth) is a key parameter that describes the asymptotic performance of many statistical methods. Numerous possibilities have been considered, and much of the literature that allows $(p, n) \to \infty$ falls into one of the following three cases — listed in order of increasing generality:

$$p^{\kappa}/n \to 0 \text{ for some exponent } \kappa \geq 1, \tag{1.1}$$

$$p/n \to c \text{ for some constant } c \geq 0, \tag{1.2}$$

$$\log(p)/n \to 0. \tag{1.3}$$

The three cases allow for increasingly large values of p relative to n, in the sense that (1.1) $\Longrightarrow$ (1.2) $\Longrightarrow$ (1.3). Much of the early work on extending M-estimation beyond fixed-p asymptotics falls into the first case [Hub73; Por84; Mam89]. The second case captures a substantial portion of the random matrix theory literature — usually with $c > 0$ [BS10]. With regard to the third case, much of the work on structured high-dimensional estimation and classification has shown that the condition $\log(p)/n \to 0$ is important in a wide variety of problems, ranging from linear regression [Neg+12], to estimation of covariance matrices and Gaussian graphical models [MB06; BL08].

Although the issue of choosing a proper scaling for p relative to n is sometimes unavoidable, another common approach is to sidestep the choice altogether by proving *non-asymptotic* results — which hold for arbitrary fixed values of n and p. This approach has become quite standard in the high-dimensional literature, because once a non-asymptotic result is available, it can be further evaluated under different limits of p and n, allowing for a wider range of possibilities to be considered. A second advantage of the non-asymptotic approach is that it reduces the technical complications that arise when studying *finite-dimensional* objects that converge to a limit in an *infinite-dimensional* space. Standard references on non-asymptotic techniques in high-dimensional statistics include [BLM13] and [Ver12].

Blessings and curses of dimensionality

High-dimensional data bring not only new structures and representations of information, but also "statistical phenomena" that fall outside the paradigm of classical statistics. Such phenomena are often referred to as the *blessings and curses of dimensionality.*[2]

In August of 2000, during the anniversary of Hilbert's famous 1900 address that set the course for much of 20th century mathematics, a conference titled "Mathematical Challenges

[2] This phrase can be traced to Bellman's work on control theory and optimization in the 1950's [Bel+61].

of the 21st Century" was held by the American Mathematical Society. At the conference, an address given by David Donoho was titled *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, and this address outlined how the effects of dimensionality may shape the horizon of mathematical research [Don00]. At that time, high-dimensional statistics was just beginning to emerge as a distinct subject of research. Writing now in 2015, the key aspects of the subject that were emphasized in Donoho's address are no less relevant today, and we describe some of them below.

Curses of dimensionality.

- **Sampling.** Monte Carlo simulation and numerical integration are fundamental aspects of statistical research, and their success is directly tied to the task of random sampling. Yet, due to the geometry of p-dimensional space, uniform random sampling becomes a hopeless task even for modest values of p. Specifically, to fill the unit cube $[0,1]^p$ uniformly with samples to a precision of ε, roughly $(\frac{1}{\varepsilon})^p$ samples are needed. As a result, one of the basic challenges of developing new methodology in high-dimensions is to modify traditional sampling-based approaches.

- **Computation.** Algorithms for numerical linear algebra and optimization lie at the core of statistical methodology. To this extent, computational barriers to implementing these algorithms are also statistical barriers. If a set of n observations in $\mathbb{R}^p$ is represented as a matrix in $\mathbb{R}^{n\times p}$ with $p/n \asymp 1$, then many basic algorithms such as matrix multiplication, matrix inversion, singular value decompositions, and so on, have complexity that is quadratic or cubic in p. Even when p is relatively tame by modern standards, e.g. $p = 10^6$, naive implementations of such algorithms are off the table. Given that the basic algorithms of applied mathematics have been highly refined over the last century, it is often sensible to use a fast but statistically suboptimal method, rather than to further refine the computational aspects of one that is statistically optimal. For this reason, the problem of identifying the right trade-off between computational cost and statistical efficiency is now at the forefront of research in high-dimensional statistics [CJ13].

- **Unidentifiable models.** As the number of parameters in a model becomes large, there is more opportunity for distinct values of parameters to specify the same distribution — resulting in a model that is *unidentifiable*. For instance, in linear regression models involving a design matrix $X \in \mathbb{R}^{n\times p}$, the distribution of observations $y \sim N(X\beta, \sigma^2)$ fails to be uniquely specified by β when $p > n$. The standard remedy for unidentifiable models in high-dimensional problems is to impose additional structural assumptions on the parameters so that identifiability is restored. Consequently, an essential challenge of high-dimensional modeling is to find forms of structure that are not overly restrictive, and yet also lead to uniquely parameterized data-generating distributions.

- **Multiple testing.** When confronted with a large number of parameters, we are often interested in discovering which parameters are "interesting", e.g. the few genes among thousands that are most strongly associated with a biological condition. If the task of assessing the importance of a single parameter is formulated as a hypothesis test, then we are dealing with a *multiple testing problem.* The basic issue underlying such problems is that if we perform p independent tests each at level $\alpha \in (0,1)$, then the chance of finding at least one false positive is at least $1-(1-\alpha)^p \approx p\alpha$, when α is small. Consequently, if we are interested in testing the significance of 1,000 genes, each one must be tested at level $\alpha = 0.00005$ just to keep the chance of a single false positive from exceeding about 5%. However, the amount of data needed to achieve even a modest degree of power at such values of α is often beyond reach. This dilemma has led researchers to focus instead on controlling the number of false positives as a fraction of the total number of rejected hypotheses (i.e. the *false discovery rate* [BH95]), rather than controlling the chance of just a single false positive. In turn, the challenges involved with controlling the false discovery rate has led to a substantial development of new methodology in high-dimensional problems.

Blessings of dimensionality.

- **Concentration of measure.** The concentration of measure phenomenon refers to the fact that a "nice" function of a large number of independent random variables is very likely to take values near its mean — provided that the variables have "light tails". An important example is the so-called *Chernoff bound* for sums of independent *sub-Gaussian* variables [BLM13]. A random variable X with mean μ is said to be sub-Gaussian with parameter σ if its moment generating function is bounded by that of a Gaussian variable with variance σ^2, i.e.

  $$\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp(\lambda^2\sigma^2/2) \quad \text{for all} \quad \lambda \in \mathbb{R}. \tag{1.4}$$

  The Chernoff bound implies that if $V_1, \ldots, V_p$ are i.i.d. sub-Gaussian variables with parameter σ, then

  $$\mathbb{P}\Big(\big|\tfrac{1}{p}\textstyle\sum_{i=1}^p V_i - \mathbb{E}[V_1]\big| > t\Big) \leq 2\exp(-\tfrac{pt^2}{2\sigma^2}). \tag{1.5}$$

  Hence, as the dimension p becomes large, the mean of the entries of $(V_1, \ldots, V_p)$ becomes extremely concentrated around its mean. More generally, the same principle can be applied to many other functions of $(V_1, \ldots, V_p)$. In this way, the concentration of measure can greatly simplify the analysis of complicated random variables by effectively reducing them to constants. The books [Led05] and [BLM13] describe many of the modern developments in probability and statistics that are connected to the concentration of measure phenomenon.

- **Universality.** The term "universality" describes situations where a sequence of random objects converges in distribution to a limit that is essentially invariant to the choice of the model that generated the objects. The statistical importance of this idea is that we may hope to understand the limiting behavior of certain quantities without assuming too much about the underlying model. Perhaps the simplest example is the ordinary central limit theorem, which shows that any standardized sum of i.i.d. random variables converges to a Gaussian limit, as long as the generating distribution has a finite second moment.

With regard to high-dimensional statistics, the significance of universality is most apparent in the context of random matrix theory. A classical example arises from the spectra of *Wigner matrices*, i.e. symmetric random matrices whose entries above the diagonal are i.i.d. variables in $L^{2+\varepsilon}$. As the dimension of such matrices diverges, a famous result of Wigner shows that their empirical spectral distribution converges to the *semi-circle law* on the real line — without any other constraint on the entry-wise distribution [Wig58]. More generally, many other instances of universality arise in statistics derived from covariance matrices, as summarized in the book of Bai and Silverstein [BS10].

- **Approach to the continuum.** In many applications, data vectors in $\mathbb{R}^p$ are actually discretized versions of continuous (or smooth) functions. Basic examples include pixelated images, or discrete time series. When p becomes large and the resolution of the observations increases, the underlying continuous structure tends to become more apparent. For instance, a discrete time series of Gaussian observations may be well-approximated by Brownian motion when the grid of time points is sufficiently fine-grained, and as a result, a large collection of analytical techniques may be brought to bear. Alternatively, when p-dimensional observations have an underlying structure involving smooth functions, a variety of tools from *functional data analysis* may be applicable [SR05].

Structure, regularization, and optimization

At first sight, the curses of high-dimensional models might seem to be inescapable facts of life. However, for many models, their *apparent* dimension is much higher than their *effective* dimension. For instance, in the context of genomics, potentially thousands of genes can be used to model a biological process, but it is well known that many processes are regulated by a specialized pathway involving a relatively small number of genes. Hence, the relevant statistical model may have an effective dimension equal to the number of genes in the pathway, whereas the apparent dimension is equal to the number of genes in the entire genome. More generally, when a high-dimensional model is controlled by a small number of "effective parameters", it is said to exhibit *low-dimensional structure*. In essence, the unifying principle of high-dimensional statistics is that a problem is tractable as long as it has low-dimensional structure. Likewise, the crucial issues in tackling many high-dimensional problems are finding this structure and using it to reduce the apparent dimension of the parameter space.

One of the most general and successful ways of harnessing low-dimensional structure is through the algorithmic tools of optimization. Given that so many statistical methods are based on the minimization of an objective function (e.g. a negative log-likelihood), it is quite natural to enforce the existence of structure by adding a *penalty function* or *regularizer*. Conceptually, the negative log-likelihood function measures lack of fit, and the regularizer measures lack of structure. Hence, by minimizing the sum of two such functions, a balance is struck between these competing forces. In more classical language, the use of a regularizer can be thought of as specifying a bias-variance tradeoff, where greater influence of the regularizer typically corresponds to more bias and less variance.

Whereas the considerations of goodness of fit and model structure are purely statistical, the choice of a regularizer is also dictated by computational cost. In fact, much of the art of high-dimensional methodology lies in finding regularizers that simultaneously balance all three of these issues. Because desired model structures can often only be described exactly in terms of non-convex regularizers (which typically lead to intractable minimization problems), a fundamental approach to finding this balance is through *convex relaxation* [BV04]. This involves replacing a non-convex penalty function with a convex "surrogate" that enforces similar structural properties. We now give some examples to illustrate this general approach.

Sparse linear regression. Consider the standard linear model involving n observations $y = (y_1, \dots, y_n)$ generated according to

$$y = X\beta + \varepsilon, \tag{1.6}$$

with Gaussian noise $\varepsilon \sim N(0, I_{p\times p})$, a deterministic design matrix $X \in \mathbb{R}^{n\times p}$ and an unknown vector of coefficients $\beta \in \mathbb{R}^p$. If the number of variables p is large, then it is often sensible to assume that many of them will have little influence on the observations. In terms of the regression coefficients, this can be formalized by supposing that the β vector is nearly sparse, i.e. many of its entries are negligible in comparison to a few relatively large entries. Hence, it is natural to penalize the number of non-zero coordinates of β, which leads to the following penalized maximum likelihood procedure,

$$\underset{v\in\mathbb{R}^p}{\text{minimize}} \quad \|y - Xv\|_2^2 + \lambda\|v\|_0, \tag{1.7}$$

where $\|v\|_0 := \#\{j \in \{1, \dots, p\} : v_j \neq 0\}$ is the so-called ℓ_0 norm and $\lambda \geq 0$ is a *regularization parameter* that controls how much weight is placed on the penalty. Unfortunately, this optimization problem is non-convex, and so a natural convex relaxation is to replace $\|v\|_0$ with $\|v\|_1$, which is the "closest" convex function to $\|v\|_0$ within the family of ℓ_q norms, i.e.

$$\underset{v\in\mathbb{R}^p}{\text{minimize}} \quad \|y - Xv\|_2^2 + \lambda\|v\|_1. \tag{1.8}$$

This procedure is known as the Lasso [Tib96], and since its inception in 1996, it has been one of the most influential methods in the subject of high-dimensional statistics. (A closely related method known as Basis Pursuit was proposed around the same time by [CDS98].) At first sight, it is not clear that the this heuristic substitution of the ℓ_1 norm will produce solutions that accurately estimate β when it is sparse. Over the course of the last 10 years, a huge wave of research developed around this question, and the conditions needed for the Lasso to succeed are now largely understood. An overview of this large theoretical development is presented in the recent book [BG11], which summarizes an enormous range of papers.

Low-rank matrix completion. Many problems involving the estimation (or completion) of an unknown matrix can be cast in the following observation model. Suppose $M \in \mathbb{R}^{d_1 \times d_2}$ is a fixed unknown matrix, and we observe

$$y_i = \langle X_i, M \rangle + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1.9}$$

where ε_i are centered i.i.d. noise variables, the $X_i \in \mathbb{R}^{d_1 \times d_2}$ are fixed and observable matrices, and $\langle A, B \rangle := \mathrm{tr}(A^\top B)$ denotes the matrix inner product. For instance, in the case of recommender systems, if our observations consist of n preferences, i.e. values M_{jk} measuring the affinity of customer j for product k, then each matrix X_i contains all 0's except for a 1 in its (j, k) entry, and the noise variables are identically 0.

When n is small compared to the dimensions d_1 and d_2, the effect of dimension is especially acute in estimating M, since the apparent number of unknown parameters is $d_1 d_2$. Clearly, some form of structure in the matrix M is necessary in order to reliably estimate M, and in many problems, it is plausible to suppose that M is nearly low rank. Returning to the example of recommender systems, it is known that many customers will have similar preferences, which is to say that many rows of M will be nearly parallel, or equivalently, that the“effective rank” of M is small. Hence, if we assume for simplicity that the noise variables are Gaussian, then a rank-penalized form of maximum likelihood estimation can be solved via the problem,

$$\underset{W \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \quad \sum_{i=1}^{n} \big(y_i - \langle X_i, W \rangle\big)^2 + \lambda \, \mathrm{rank}(W), \tag{1.10}$$

where $\lambda \geq 0$. In parallel with the problem of sparse linear regression above, the rank penalty leads to an intractable combinatorial optimization problem. Also, the standard relaxation for the rank-regularized problem is strikingly similar to the ℓ_1 relaxation for the ℓ_0 problem. To see the connection, if we let $\sigma(W)$ denote the vector of singular values of the matrix W, then note that $\mathrm{rank}(W) = \|\sigma(W)\|_0$. Hence, the success of the ℓ_1-heuristic for ℓ_0 minimization leads us to consider using $\|\sigma(W)\|_1$ as a regularizer in the matrix context. Specifically, the ℓ_1 norm of the singular values is more commonly referred to as the nuclear norm (or Schatten-1 norm), denoted $\|W\|_*$, which yields the following convex relaxation of (1.11),

$$\underset{W \in \mathbb{R}^{d_1 \times d_2}}{\text{minimize}} \quad \sum_{i=1}^{n} \big(y_i - \langle X_i, W \rangle\big)^2 + \lambda \, \|W\|_*. \tag{1.11}$$

The statistical properties of this procedure, as well as many others based on nuclear norm minimization, have been intensively studied over the last few years. See for instance the papers [KLT11], [RFP10], [CR09], [CP11], [NW12].

1.2 Motivations of the dissertation

This dissertation is motivated by an interest in two facets of the theory of high-dimensional statistics: *inference* and *structure.*

Inference. When using the word "inference", we have in mind problems involving confidence intervals, hypothesis tests, uncertainty quantification, and so on. The common thread among such problems is that they are concerned with measuring the fluctuations of a statistic, or finding its limiting distribution.

While the past decade has seen tremendous progress in the study of high-dimensional models, relatively little attention has been given to problems of inference, and only recently has substantial interest shifted in this direction. The trend has been particularly strong with regard to significance testing in high-dimensional linear models, as evidenced by the recent works [ZZ14; JM14a; JM14b; Buh13; Van+14]. As work in this direction continues, there remain many open questions regarding the effects of high dimension on traditional inference tools, and a complete understanding of high-dimensional inference at large seems to lie in the distant future.

As a starting point towards a more general theory of high-dimensional inference, it is natural to investigate *bootstrap methods*, since they provide a unified framework for many confidence intervals and hypothesis tests. To this end, Chapter 2 of the dissertation will consider the *residual bootstrap* method in the context of the high-dimensional linear model. There, our focus will be in understanding how the structure of the design matrix affects the performance of the residual bootstrap in approximating the sampling distribution of linear contrasts. Bootstrap methods are reviewed in Section 1.3.

Structure. As described in Section 1.1, low-dimensional structure is often the crucial ingredient needed for statistical methods to work in high-dimensional problems. Likewise, the field of high-dimensional statistics has been influenced by the viewpoint that the existence of low-dimensional structure should be taken as a default assumption — because in the absence of structure, it is unlikely that much progress can be made. This line of thought has been codified in the so-called *bet on sparsity principle* [HTF09, Section 16.2.2]. Specifically, in models where sparsity is the relevant form of structure, the principle asserts that it is best to "use a procedure that does well in sparse problems, since no procedure does well in dense problems".[3]

[3] It is not our intention here to criticize the bet on sparsity principle. We only mention it as a way of giving some context to the general problem of checking structural assumptions in high-dimensional models.

While the practice of betting on structured models has led to great advances in methodology, this success may sometimes overshadow more a basic question: How do we know if it is necessary to bet on structure? Perhaps surprisingly, the answer is not always clear. In this regard, the high-dimensional literature has not given much consideration to the possibility that some structural assumptions can be verified in a data-driven way. Chapter 3 of this dissertation is motivated by a desire to investigate this issue in greater detail, and to identify situations where the existence of low-dimensional structure can be confirmed empirically. Given that sparsity in the linear model is a canonical example of low-dimensional structure, it will serve as our cornerstone. In particular, the chapter develops a method to estimate the "effective number" of non-zero coefficients in a high-dimensional linear model — without relying on any sparsity assumptions. Because the method depends on the ability to randomly generate the design matrix in a special way, the method is geared primarily to the framework of compressed sensing, where such control can be arranged by way of a physical measurement system. The topic of compressed sensing is reviewed in Section 1.4.

Relations between inference and structure. Inference and structure have a chicken-and-egg relationship in high-dimensional statistics. On one hand, hypothesis tests and confidence intervals are naturally suited to the task of confirming model structure. On the other hand, many of these procedures will not work properly in high dimensions if structure is not available.

One way of disentangling the chicken-and-egg dilemma is to play different forms of structure off of one another. This idea is especially useful when one of the structures is relatively easy to verify. For instance, in linear models, structure in the (observed) design matrix tends to be easier to verify than structure in the (unknown) regression coefficients. Indeed, it is a basic theme of the dissertation that structured designs can obviate the need for sparsity in the regression coefficients when doing inference. In this regard, we show in Chapter 2 that when the design matrix is nearly low rank, it is possible to do inference on linear contrasts without making any sparsity assumptions. Secondly, in Chapter 3, we show that when the design matrix can be sampled in a prescribed way, it is possible to "test the hypothesis of sparsity", again without making any sparsity assumptions.

1.3 Bootstrap methods

Questions of inference often boil down to finding the sampling distributions of estimators and test statistics. Because analytical approximations to such distributions are rarely available outside of specialized models, a more general approach is often needed. *Resampling methods* are designed to serve this purpose.

In order to approximate the sampling distribution of a statistic, it is necessary to measure how much a statistic varies over repeated experiments. When the outcome of only a single experiment is available, resampling methods attempt to circumvent this difficulty by evaluating the statistic on certain subsets of the full dataset. The values obtained on subsets then serve as "proxies" for values of the statistic in repeated experiments. Since this is such an intuitive idea, it is difficult to precisely trace its historical origin. Even so, the earliest work on resampling is commonly attributed to Quenouille [Que49], and Tukey [Tuk58], who are credited with the *jackknife* method.[4]

The jackknife method involves splitting a sample of n observations into n distinct subsets of size $n-1$, by excluding one observation at a time. (Generalizations to subsets of other sizes are also possible.) The work of Quenouille and Tukey showed that by aggregating the values of a statistic over such subsets, it is possible to correct for bias and construct sensible confidence intervals in certain problems. During the 1960's and 1970's, these early ideas surrounding the jackknife were extended further, as in the work of Hartigan [Har69; Har75] and Stone [Sto74]. In turn, during the late 1970's, the understanding of resampling methods was profoundly reshaped by the work of Efron [Efr79], who introduced *bootstrap* methods, and showed that the jackknife can be viewed as an approximation thereof.[5]

At a high level, the "bootstrap principle" can be explained along the following lines. Suppose observations $X_1,\dots,X_n$ are drawn in an i.i.d. manner from a distribution F. Also, let $T = g(X_1,\dots,X_n)$ denote a statistic of interest, where g is a fixed function. If the distribution F were known to us, then we could generate an independent copy of the original dataset, $X_1',\dots,X_n'$, and then compute an additional sample of the statistic using $g(X_1',\dots,X_n')$. By repetition, we could then obtain an unlimited number of samples of T, yielding the desired sampling distribution. The fundamental idea of the bootstrap is to carry out this scheme by replacing F with the empirical distribution $\widehat{F}$ of the original sample. Said differently, this involves generating many i.i.d. samples of size n according to $X_1^*,\dots,X_n^* \sim \widehat{F}$, and then computing $T^* := g(X_1^*,\dots,X_n^*)$. Repeatedly generating samples of T^* in this way yields the so-called *non-parametric bootstrap* approximation to the desired sampling distribution of T.

Although the bootstrap is simple to describe conceptually, the mathematical problem of understanding when it works, as well as the methodological problem of tailoring it to specific situations, have spawned more than three decades of research on a large family of "bootstraps". One natural variation on the non-parametric bootstrap is to sample datasets $X_1^*,\dots,X_m^*$ of size m rather than n, where $m \ll n$. When the sampling is done with replacement, this is typically referred to as the *m-out-of-n* bootstrap [BGZ97], and when it is done without replacement, the method is referred to as *subsampling* [PRW99]. These strategies can lead to advantages over the non-parametric bootstrap when dealing with "non-smooth" statistics or heavy-tailed data (see Section 2.3 of the book [PRW99] for detailed examples). Alternatively, when it is known that the data-generating distribution F belongs to a parametric family, say $F = F_\theta$, then this information can be exploited by drawing i.i.d. samples $X_1^*,\dots,X_n^* \sim F_{\widehat{\theta}}$, where $\widehat{\theta}$ is an estimate obtained from the original dataset. This recipe is known as the *parametric bootstrap.* Lastly, with regard to linear regression models, statistics arising from estimated coefficients can be approximated in law via the *residual*

[4] In the preface of the book [PRW99], it is remarked that an even earlier manifestation of resampling occurs in the work of Mahalanobis [Mah46].

[5] A modern survey of the jackknife and the bootstrap can be found in the book of [ST95].

bootstrap. Since this method will form the basis of our work in Chapter 2, we now review it in detail.

The residual bootstrap

In the setting of the standard linear regression model with fixed design, many statistics of interest are functionals of the estimated regression coefficients. The residual bootstrap (RB), proposed by Efron in 1979 [Efr79], is one of the most widely used resampling strategies for approximating the laws of such statistics.

To introduce the method, suppose we observe a set of values $y = (y_1, \dots, y_n)$ generated from the model

$$y = X\beta + \varepsilon, \tag{1.12}$$

where $X \in \mathbb{R}^{n\times p}$ is a fixed design matrix with full rank and $p < n$. The coefficient vector $\beta \in \mathbb{R}^p$ is unknown, and the entries of $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ are drawn i.i.d. according to an unknown distribution F_0 with mean 0, and unknown variance $\sigma^2 < \infty$. Also, let

$$\widehat{\beta}_{\text{LS}} := (X^\top X)^{-1} X^\top y \tag{1.13}$$

denote the least squares estimator, and suppose we are interested in approximating the distribution of $g(\widehat{\beta}_{\text{LS}})$, where $g : \mathbb{R}^p \to \mathbb{R}$ is a smooth function. Clearly, the randomness in $\widehat{\beta}_{\text{LS}}$ is generated entirely from ε. Hence, it is natural to design a resampling scheme that is based on generating "approximate samples" of ε. If we denote the residuals of $\widehat{\beta}_{\text{LS}}$ by $\widehat{\varepsilon} := y - X\widehat{\beta}_{\text{LS}}$, then the algebraic relation

$$y = X\widehat{\beta}_{\text{LS}} + \widehat{\varepsilon} \tag{1.14}$$

makes it is plausible that $\widehat{\varepsilon}$ might behave like a genuine sample of ε. To make this idea more precise, we will use the residuals to construct an approximation to F_0. Specifically, consider the distribution $\tilde{F}_n$ that places mass $1/n$ at each of the values $\widehat{\varepsilon}_i - \bar{e}$, where $\bar{e} := \frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i$. (Subtracting off the value $\bar{e}$ merely ensures that $\tilde{F}_n$ has mean 0, as F_0 does.) With the distribution $\tilde{F}_n$ in hand, we are free to generate a limitless supply of vectors

$$\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*) \overset{\text{i.i.d.}}{\sim} \tilde{F}_n. \tag{1.15}$$

Likewise, if we use $\widehat{\beta}_{\text{LS}}$ as a proxy for β, then we can generate a proxy for y according to

$$y^* := X\widehat{\beta}_{\text{LS}} + \varepsilon^*. \tag{1.16}$$

In turn, a least-squares estimator computed from y^* will be denoted by $\widehat{\beta}^*_{\text{LS}}$. That is,

$$\widehat{\beta}^*_{\text{LS}} := (X^\top X)^{-1} X^\top y^*. \tag{1.17}$$

In this notation, the RB approximation to the law of $g(\widehat{\beta}_{\text{LS}})$ is simply a histogram constructed from a large number of samples of $g(\widehat{\beta}^*_{\text{LS}})$.

Consistency and failure of the residual bootstrap

The groundwork for the asymptotic theory of the RB method was laid in two seminal papers of Bickel and Freedman [Fre81; BF83]. In essence, these papers show that if $p/n \to 0$, then the RB method consistently approximates the laws of all contrasts involving least-squares coefficients. Moreover, the condition $p/n \to 0$ is necessary to the extent that if $p/n \to \kappa \in (0,1)$, then there is always at least one contrast for which the RB method fails (when applied to least-squares). We will focus on the main results of the 1983 paper, since they partly subsume those of the 1981 paper. Both papers deal with the standard regression model with fixed design (1.12), under the assumptions stated in Section 1.3.[6] In summarizing these results with $(n,p) \to \infty$, we will view the regression model (1.12) as being embedded in a sequence of models where p, β, and X depend implicitly on n, but where F_0 is fixed with respect to n.

Probability metrics. The results describing the consistency and failure of the RB method will make use of two metrics on probability distributions: the *Mallow-ℓ_2 metric*[7] and the *Lévy-Prohorov metric.* For any two random vectors U and V in a Euclidean space with ℓ_2 norm $\|\cdot\|_2$, the associated Mallows metric d_2 is defined according to

$$d_2^2(\mathcal{L}(U), \mathcal{L}(V)) := \inf_{\pi \in \Pi} \Big\{ \mathbb{E}\Big[\|U - V\|_2^2 \Big] : (U, V) \sim \pi \Big\}, \tag{1.18}$$

where $\mathcal{L}(\cdot)$ denotes the law of a random variable, and the infimum is over the class Π of joint distributions π whose marginals are $\mathcal{L}(U)$ and $\mathcal{L}(V)$. It is worth noting that convergence in d_2 is strictly stronger than weak convergence, since it also requires convergence of second moments. Additional properties of the metric are discussed in the paper [BF81].

Next, if μ and ν are two distributions on $\mathbb{R}^d$, then Lévy-Prohorov metric d_{LP} is defined as

$$d_{\text{LP}}(\mu,\nu) := \inf \Big\{ \epsilon > 0 : \mu(K) \leq \nu(K^\epsilon) + \epsilon \text{ and } \nu(K) \leq \mu(K^\epsilon) + \epsilon \text{ for all } K \in \mathcal{K}(\mathbb{R}^d) \Big\}, \tag{1.19}$$

where $\mathcal{K}(\mathbb{R}^d)$ is the collection of compact subsets of $\mathbb{R}^d$, and K^ϵ is the set of points in $\mathbb{R}^d$ whose Euclidean distance to K is at most ϵ. The d_{LP} metric has the important property that it metrizes weak convergence of distributions on $\mathbb{R}^d$ (and in fact separable metric spaces). It can also be shown that $d_{\text{LP}}(\mu,\nu) \leq d_2(\mu,\nu)^{2/3}$ for all μ and ν [BF83]. We refer to the paper [GS02] for further details on the d_{LP} metric and its relations with other metrics.

Consistency. The problem of interest is to approximate the law of a general contrast involving least-squares coefficients,

[6] The paper [Fre81] considers random design as well.

[7] The Mallows metric has many other names in the literature, such as the Wasserstein metric, the Kantorovich metric, and the earth-mover distance. Also, the d_2 metric can be generalized by allowing any ℓ_q norm with $q \geq 1$ to replace the ℓ_2 norm in definition (1.18).

$$\Psi(F_0;c) := \mathcal{L}(c^\top(\widehat{\beta}_{\text{LS}} - \beta)), \tag{1.20}$$

where $c \in \mathbb{R}^p \setminus \{0\}$ is arbitrary. If we let $\tilde{F}_n$ denote the empirical law of the centered least-squares residuals, as in Section 1.3, then the law of the RB approximation, conditionally on y, is given by

$$\Psi(\tilde{F}_n;c) = c^\top(\widehat{\beta}^*_{\text{LS}} - \widehat{\beta}_{\text{LS}}). \tag{1.21}$$

In order to measure how well the random distribution $\Psi(\tilde{F}_n;c)$ approximates $\Psi(F_0;c)$, the Mallows-ℓ_2 metric will be used. As a way of comparing $\Psi(F_0;c)$ and $\Psi(\tilde{F}_n;c)$ on the proper scale, we will normalize by the standard deviation of $\Psi(F_0;c)$. A simple calculation shows that the variance of this distribution is

$$v(X;c) := \text{var}(\Psi(F_0;c)) = \sigma^2 c^\top (X^\top X)^{-1} c, \tag{1.22}$$

and we will write v instead of $v(X;c)$ to lighten notation. Under this normalization, the following theorem shows that RB consistency is uniform with respect to the choice of contrast, provided that p is small with respect to n.

Theorem 1.1 ([BF83])**.** *Suppose the assumptions of the model* (1.12) *hold, and that as* $(n,p) \to \infty$*, we have* $p/n \to 0$*. Then as* $(n,p) \to \infty$*,*

$$\mathbb{E}\left[\sup_{c \in \mathbb{R}^p \setminus \{0\}} d_2^2\Big(\tfrac{1}{\sqrt{v}}\Psi(\tilde{F}_n;c), \tfrac{1}{\sqrt{v}}\Psi(F_0;c)\Big)\right] \to 0. \tag{1.23}$$

Failure. Under the condition $p/n \to \kappa \in (0,1)$, the paper [BF83] demonstrates the failure of RB approximation by way of specific contrasts. The construction of "bad" contrasts for which RB fails is closely connected to the structure of the design matrix, and its associated "hat matrix"

$$H := X(X^\top X)^{-1}X^\top. \tag{1.24}$$

The ith diagonal entry H_{ii} (also known as a leverage score) measures the statistical leverage of the ith design point X_i with respect to the fitted coefficients $\widehat{\beta}_{\text{LS}}$. It is a basic fact that for any design matrix, the leverage scores satisfy

$$\max_{1 \le i \le n} H_{ii} \ge p/n, \tag{1.25}$$

which shows that when $p/n \to \kappa \in (0,1)$, there is always at least one design point with non-negligible leverage. The contrasts used to demonstrate the failure of RB are based on the existence of such design points in the high-dimensional setting.

To simplify the presentation of the counterexamples, define λ_n to be the distribution that places mass $1/n$ at each of the numbers $1 - H_{ii}$, with $i = 1, \dots, n$. Since the distributions λ_n are defined on the compact set $[0,1]$, they form a tight sequence, and by Prohorov's theorem, there is a subsequence along which λ_n converges to a weak limit, say λ. The choice of contrasts for demonstrating the failure of RB depends on whether or not the distribution λ is degenerate (i.e. point mass at the value $\int t d\lambda(t) = 1 - \kappa$).

Theorem 1.2 ([BF83][8])**.** *Assume the model* (1.12) *holds and there is a constant κ such that as $(n, p) \to \infty$, we have $p/n \to \kappa \in (0, 1)$. Then the statements below are true.*

(i) Suppose the distribution λ is nondegenerate, and the noise distribution F_0 is equal to $N(0, 1)$. Let $c^\top = X_{i^\star}^\top$ be the row of X corresponding to $i^\star = \text{argmax}_{1\leq i\leq n} H_{ii}$. Then, the sequence of random variables $d_{\text{LP}}(\frac{1}{\sqrt{v}}\Psi(\tilde{F}_n; c), \frac{1}{\sqrt{v}}\Psi(F_0; c))$ does not converge to 0 in probability.

(ii) Suppose the distribution λ is degenerate. In this case, there is at least one subsequence $\{i_n\}$ along which $H_{i_n i_n} \to \kappa$. Let $c^\top$ be the i_nth row of the design along such a subsequence, i.e. $c^\top = X_{i_n}^\top$. Also suppose the noise distribution F_0 is symmetric about 0, has a finite moment-generating function in an open interval about 0, and cannot be represented as a convolution of $(1-\kappa)F_0$ with another distribution. Then, the sequence of random variables $d_{\text{LP}}(\frac{1}{\sqrt{v}}\Psi(\tilde{F}_n; c), \frac{1}{\sqrt{v}}\Psi(F_0; c))$ does not converge to 0 in probability.

The question of finding a remedy for these counterexamples will be the subject of Chapter 2.

1.4 Compressed sensing

Historically, signal processing research has treated data acquisition and data compression as distinct processes. Over time, this division has created a situation where the tools for acquisition and compression are at cross purposes. In one direction, acquisition devices have made it possible to sample signals at higher frequencies — producing larger amounts of data at finer resolution. Meanwhile, in the opposite direction, the methods of data compression have found new ways to faithfully represent complex signals while retaining smaller amounts of the acquired data. These conflicting aims have led researchers to question whether or not the division is really necessary, and to seek more efficient approaches. Indeed, to use a well quoted remark of Donoho, “Why go to so much effort to acquire all the data when most of what we get will be thrown away? Can’t we just directly measure the part that won’t end up being thrown away?” [Don06]. In response to questions such as these, *compressed sensing* (CS) has emerged as an alternative signal processing framework that integrates acquisition and compression into a single process.[9]

To make the relationship between CS and conventional signal processing more concrete, consider the following stylized description of an imaging problem. In order to obtain a high quality image, a typical camera first acquires several million pixels of data from a light source. In turn, the megabytes of raw pixels might be compressed into 100 kilobytes of data in a specialized format for later retrieval. By contrast, a CS camera would acquire only 100 kilobytes of raw data at the outset. Also, instead of collecting the data in terms of

[8] We have reformulated the result slightly from its original form.

[9] This idea is reflected in the the phrase compressed sensing, since the word “sensing” is roughly synonymous with “data acquisition”.

ordinary pixels, a CS camera would collect "random linear combinations of pixels," which can be transformed into a 100 kilobyte image.[10] Although both cameras are able to produce high quality images, the fact that the CS camera demands far less data acquisition turns out to be a substantial practical advantage in certain situations. For instance, when the light source falls outside of the visible spectrum, e.g. in infrared or terahertz imaging, light sensors with large numbers of pixels can be very expensive to build, and moreover, the rate of acquisition can be very slow [Cha+08]. Consequently, the ability of a CS camera to reduce data acquisition can lead to a reduction in cost, and an increase in speed. Beyond the context of optical imaging, other applications of compressed sensing are being actively studied in numerous areas, such as magnetic resonance imaging, radar, and cognitive radio, just to name a few [EK12; FR13].

As sensible as the premise of CS may appear from the previous example, the early research in CS was initially met with some degree of skepticism [Mac09]. The skepticism was rooted in certain ideas surrounding the *Shannon-Nyquist sampling theorem*, which is a foundational result in classical sampling theory [Nyq28; Sha49]. The theorem addresses the problem of reconstructing a bandlimited continuous-time signal[11] via uniform sampling. Roughly speaking, the theorem asserts that if the signal's bandwidth is at most B, then it is possible to exactly reconstruct the signal from $2B$ uniformly spaced samples in the time domain. In this way, if we imagine a continuous-time signal as being represented by a vector in $\mathbb{R}^p$, with coordinates corresponding to frequency components of the continuous signal's (discretized) Fourier transform, then one might expect that a reliable reconstruction would require the number of measurements to increase proportionally with p. On the other hand, the premise of CS is that if a signal in $\mathbb{R}^p$ can be faithfully compressed into a vector in $\mathbb{R}^k$ with $k \ll p$, then it should be possible to accurately reconstruct the signal using a number of measurements that is of order k. Despite the perceived conflict between these two ideas, there is in fact no logical contradiction. In essence, the Shannon-Nyquist theorem is a "worst-case" result that provides a sufficient condition for recovery, but not a necessary one. Meanwhile, the focus of CS is restricted to a special class of "compressible signals" for which the classical theory is overly pessimistic. So, in essence, CS seeks to overcome the curses of dimensionality by making use of compressibility.

Signal compressibility not only makes efficient data acquisition possible, but it is also ubiquitous in nature. Indeed, it is often the case that natural signals have a *sparse representation* with respect to a specialized basis or "dictionary". Perhaps the most well known types of bases and dictionaries used in signal processing are based on *wavelets*, and the design of carefully engineered families of wavelets for specific problems in image and audio processing has become an extensive field of research [Mal08]. In detail, a signal $x \in \mathbb{R}^p$ is said to have a sparse representation with respect to a dictionary $\mathcal{D} = \{\phi_1, \dots, \phi_p\} \subset \mathbb{R}^p$ if there exists an expansion $x = \sum_{i=1}^p c_i\phi_i$ for which most of the coefficients c_i are small. If the coefficients are sorted in order of decreasing magnitude, $|c_1| \geq |c_2| \geq \dots \geq |c_p|$, and if $|c_i| \approx 0$ for all i

[10] An example of a camera constructed along these lines is the so-called "single-pixel camera" [Dua+08].

[11] A continuous-time real-valued signal is bandlimited if its Fourier transform has bounded support.

greater than some number $T \in \{1, \dots, p\}$, then it is clear that x is well-approximated by its *T-term approximation* $\sum_{i=1}^{T} c_i \phi_i$. To the extent that the T-term approximation is a compressed version of x, a sparse representation of x has become a standard way of describing the compressibility of x. In the ideal case where $c_i = 0$ for all $i \geq T$, the signal x is called *T-sparse* with respect to $\mathcal{D}$, or is said to exhibit "hard sparsity". Although the notion of hard sparsity offers a convenient mathematical way of quantifying how compressible a signal is, hard sparsity gives an unsatisfactory answer for signals with many small (but non-zero) coordinates. In Chapter 3, we will introduce a generalized family of sparsity measures as a means of counting the "effective number" of coordinates of arbitrary signals.

Having summarized the conceptual rudiments of CS, we turn our attention in the next subsection to the mathematical aspects of the subject, including the formulation of the CS model, recovery algorithms, and theoretical results.

The formulation and theory of compressed sensing

The theoretical foundation CS is built upon the *linear measurement model.* The observations in this model, referred to as "linear measurements", are given by

$$y_i = \langle a_i, x \rangle + \sigma \epsilon_i, \qquad i = 1, \dots, n, \tag{1.26}$$

where the $a_i \in \mathbb{R}^p$ are user-specified "measurement vectors", $x \in \mathbb{R}^p$ is an unknown signal, and the $\sigma \varepsilon_i$ are noise variables with $\sigma > 0$ being a constant representing the noise level. The model is high-dimensional and underdetermined, in the sense that $n \ll p$. In matrix notation, the observations $y = (y_1, \dots, y_n)$ may be expressed more concisely as

$$y = Ax + \sigma \epsilon, \tag{1.27}$$

where $A \in \mathbb{R}^{n \times p}$ is the "measurement matrix" whose ith row is a_i, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)$. Conventionally, it is also assumed that there exists an orthonormal basis for $\mathbb{R}^p$ in which x is sparse, or nearly sparse. That is, if the relevant basis matrix is denoted by $\Phi \in \mathbb{R}^{p \times p}$, then most of the entries of the vector Φx are nearly 0. But for mathematical convenience, it is simpler to imagine that x is sparse in the standard basis of $\mathbb{R}^p$, and this can be achieved without loss of generality by absorbing Φ into A from the right, and identifying x with Φx.

Despite the resemblance of the linear measurement model (1.27) to linear regression (as discussed in Section 1.3), the role of the measurement (design) matrix is often different in these two settings. In general, the modern regression literature treats the design matrix as being "given", and the possibility of constructing the matrix in different ways to suit different statistical goals is typically given relatively little consideration.[12] By contrast, in CS, the matrix A usually corresponds to the configuration of a measurement device, and the response y represents a physical interaction between the device and the signal x. For

[12] To some extent, this belies the historical connection between the term "design matrix" and the subject of experimental design.

this reason, it is often assumed that the system designer has control over the choice of A. However, depending upon the physical limitations of certain measurement systems, there are some choices of A that, in spite of their favorable theoretical properties, are difficult to realize in actual experiments. The problem of identifying choices of A that balance the competing demands of theory and practice is an area of ongoing research, from both mathematical and experimental perspectives [DE11, Section IV].

Restricted isometries. Regardless of the particular way that the matrix A is constructed, the assumption $n \ll p$ implies that A always has a nontrivial null space in the standard CS model. In particular, this means that the model is unidentifiable in the sense that there always exists a non-zero x that cannot distinguished from the 0 vector using only the measurements y. It is in this respect that sparsity plays a crucial role in CS model by restoring identifiability. If it is assumed that x is non-zero and T-sparse in the standard basis for $\mathbb{R}^p$, then a clearly minimal requirement for identifiability is that $Av \neq 0$ for all non-zero T-sparse vectors $v \in \mathbb{R}^p$. In fact, as simple as this consideration may seem, it contains the essential idea underlying many measurement schemes.

As the subject of CS has developed, a variety of identifiability conditions have been analyzed. Two specific examples that have been especially influential in the literature are the *restricted isometry property of order k* (RIP-k) [CT05], and the *null-space property of order k* (NSP-k) [CDD09; DH01], where k is a presumed upper bound on the sparsity level of the true signal. In detail, a matrix $A \in \mathbb{R}^{n \times p}$ is said to satisfy RIP-k if there exists a number $\delta_k \in (0, 1)$ such that the bounds

$$(1 - \delta_k)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta_k)\|x\|_2 \tag{RIP-k}$$

hold for all k-sparse vectors $x \in \mathbb{R}^p$. Secondly, a matrix A is said to satisfy NSP-k if there is a constant $C_k > 0$ such that the bound

$$\|v[S]\|_1 \leq C_k \|v[S^c]\|_1 \tag{NSP-k}$$

holds for all $S \subset \{1, \dots, p\}$ with $\text{card}(S) \leq k$, and all v in the nullspace of A. Here, $v[S] \in \mathbb{R}^p$ is the vector obtained by setting the jth coordinate of v equal to 0 if $j \notin S$.

The importance of RIP-k and NSP-k is that they provide criteria that lead to provable guarantees about signal recovery, and also are satisfied by a fairly large class of measurement matrices that can be approximated in physical systems. Moreover, these properties are “generic”, in the sense that they hold with high probability for matrices drawn from suitable ensembles. As an example, the following theorem quantifies this idea for sub-Gaussian measurement matrices (recall the definition of a sub-Gaussian random variable in line (1.4)).

Theorem 1.3 ([Ver12])**.** *Let G_0 be a centered sub-Gaussian distribution on $\mathbb{R}$ with unit variance, and let $A \in \mathbb{R}^{n \times p}$ be a random matrix whose entries are drawn in an i.i.d. manner from $\frac{1}{\sqrt{n}} G_0$. Then, there is a constant c depending only on the distribution G_0 such that the following is true. If $\delta \in (0, 1)$, $k \in \{1, \dots, p\}$, and*

$$n \geq \tfrac{c}{\delta^2} k \log(ep/k), \tag{1.28}$$

then with probability at least $1 - 2\exp(-c\delta^2 n)$*, the matrix* A *satisfies* RIP-k *with* $\delta_k \leq \delta$.

A similar statement also holds for NSP-k, since it is implied by RIP-k for certain choices of C_k and δ_k (see [EK12, Theorem 1.5]).

Recovery algorithms. During the past ten years, a remarkably large number of algorithms for recovering sparse signals have been proposed and analyzed in the CS literature. While many of these algorithms have substantial technical differences, they can be broadly unified as "searching for the sparsest signal that is consistent with the data". In the simplest case of noiseless measurements where $y = Ax$, this idea can be formulated directly as

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\text{minimize}} \quad & \|v\|_0 \\ \text{subject to} \quad & y = Av. \end{aligned} \tag{1.29}$$

Beyond its intuitive appeal, the optimization problem (1.29) offers an attractive approach because it is known that if A satisfies certain identifiability conditions, then the true signal x is in fact the unique solution. For instance, this is known to be the case when x is k-sparse and A satisfies RIP-$2k$. Yet, in spite of its advantages, this approach is severely limited in two respects. First, the problem (1.29) is computationally intractable in the sense that it is NP-hard [FR13, Section 2.3]. Second, the approach has limited applicability to physical systems, since the constraint $y = Av$ does not allow for measurement noise.

In order to handle these limitations, some of the most well known recovery algorithms use the ℓ_1 norm as a convex surrogate for the ℓ_0 norm, and also relax the equality constraint in (1.29) to an approximate version. One such algorithm is Basis Pursuit Denoising (BPDN) [CDS98], which is formulated as the minimization problem

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\text{minimize}} \quad & \|v\|_1 \\ \text{subject to} \quad & \|y - Av\|_2 \leq \sigma\epsilon_0. \end{aligned} \tag{BPDN}$$

where ϵ_0 is a constant that is assumed to satisfy $\|\epsilon\|_2 \leq \epsilon_0$ for all realizations of the noise vector ϵ in the model (1.27). Since the problem (BPDN) is convex, efficient numerical solvers are available, as described in the paper [VF08] and references therein. A number of other optimization-based approaches similar to BPDN have also been proposed, such as the Lasso [Tib96] and the Dantzig Selector [CT07]. These methods are known to enjoy performance guarantees that are comparable to BPDN, but for brevity we do not describe them in detail.

In addition to methods that can be implemented as a single optimization problem, another large collection of methods are formulated as iterative or "greedy" procedures. The principal advantage of these methods is that they tend to be fast, especially when the true signal is very sparse. Roughly speaking, these algorithms proceed in the following way. If the true signal is assumed to be k-sparse, then an initial support set is chosen, say $S^0 \subset \{1, \ldots, p\}$

with $\mathrm{card}(S) \leq k$, and an initial estimate $\widehat{x}^0$ is easily computed as the restricted least-squares solution

$$\widehat{x}^0 := \operatorname{argmin}\{\|y - Av\|_2 : v \in \mathbb{R}^p \text{ and } \mathrm{supp}(v) \subset S^0\}, \tag{1.30}$$

where $\mathrm{supp}(v) \subset \{1, \dots, p\}$ denotes the set of indices j such that $v_j \neq 0$. Then, the support set S^0 is updated to a new set S^1 according to a variable selection rule involving $\widehat{x}^0$, and a new estimate $\widehat{x}^1$ is computed by replacing S^0 with S^1 in line (1.30). Examples of popular methods that fall into this framework include Orthogonal Matching Pursuit [MZ93; PRK93], Compressive Sampling Matching Pursuit (CoSamp) [NT09], and Subspace Pursuit [DM09]. These algorithms have been analyzed in depth and are known to have recovery properties similar to the Dantzig Selector, Lasso, and BPDN. A more detailed review of these methods may be found in the book [FR13].

Theoretical results. As a way of illustrating the general nature of recovery guarantees in CS, we present two fundamental results for the BPDN algorithm below. These results address both achievability and optimality, providing an upper bound on the ℓ_2 approximation error, and also revealing that for a certain class of compressible signals, the rate of approximation error cannot be improved by any recovery algorithm whatsoever.

The following upper bound on BPDN approximation error first appeared in the paper [CRT06] and was subsequently refined in later years. Our statement of the result differs slightly from its original form in order to reflect some of these refinements [CWX10]. To fix some notation, if $k \in \{1, \dots, p\}$, we define $x_{|k} \in \mathbb{R}^p$ to be the best k-term approximation to $x \in \mathbb{R}^p$, i.e. the vector obtained by setting the $p - k$ smallest entries of x (in magnitude) to 0.

Theorem 1.4 ([CRT06; CWX10]). *Suppose that the model* (1.27) *holds, and all realizations of the nose vector satisfy* $\|\epsilon\|_2 \leq \epsilon_0$ *for some absolute constant* $\epsilon_0 \geq 0$. *Fix* $k \in \{1, \dots, p\}$, *and let* A *be a matrix that satisfies* RIP-k *with* $\delta_k < 0.307$. *Then, there are absolute constants* c_1 *and* c_2 *such that any solution* $\widehat{x}$ *to* (BPDN) *satisfies*

$$\|\widehat{x} - x\|_2 \leq c_1 \sigma \epsilon_0 + c_2 \tfrac{\|x - x_{|k}\|_1}{\sqrt{k}}. \tag{1.31}$$

There are several notable aspects of the theorem. First, it applies to any signal $x \in \mathbb{R}^p$, and to any realization of the noise variables, provided that they obey the stated ℓ_2 norm bound. Second, the result generalizes some of the earliest recovery guarantees in CS, which only apply in the case of noiseless measurements and hard-sparse signals. In particular, if $\epsilon_0 = 0$ and $\|x\|_0 \leq k$, then the result guarantees *exact* recovery, $\widehat{x} = x$.

We now consider the question of whether or not the rate of approximation error in (1.31) can possibly be improved by another recovery algorithm.[13] For this purpose, we restrict our attention to signals lying in a *weak* ℓ^q ball of radius $R \geq 0$,

$$\mathcal{B}_q(R, p) := \{v \in \mathbb{R}^p : |v|_{[i]} \leq R \cdot i^{-1/q}\}, \tag{1.32}$$

[13] Our discussion here follows the presentation of ideas in Sections 3.3 and 3.5 of the paper [Can06].

where $|v|_{[1]} \geq \cdots \geq |v|_{[p]}$ are the sorted magnitudes of the coordinates of a vector v. This class of signals is compressible in the sense that any signal $x \in \mathcal{B}_q(R,p)$ is well approximated by its best k-term approximation for sufficiently large k. Specifically, if $x \in \mathcal{B}_q(R,p)$, then it is known that for any $k \in \{1,\dots,p\}$,

$$\|x - x_{|k}\|_1 \lesssim R \cdot k^{1-1/q}. \tag{1.33}$$

Details may be found in the paper [CDD09]. Consequently, in the case of noiseless measurements, if we apply Theorem (1.4) with a suitable matrix A, and use line (1.33) with the choice $k = \lceil n/\log(ep/n) \rceil$, then for any $x \in \mathcal{B}_q(R,p)$ with $q \in (0,1)$, we have the following bound for all solutions $\widehat{x}$ of (BPDN),

$$\|\widehat{x} - x\|_2 \lesssim R \cdot \left(\tfrac{\log(ep/n)}{n}\right)^{1/q-1/2}. \tag{1.34}$$

The significance of this bound is that it can be directly linked to a universal lower bound on the optimal ℓ_2 approximation error for recovering signals in $\mathcal{B}_q(R,p)$ from linear measurements. To be precise, define the minimax ℓ_2 approximation error

$$\mathcal{M}_q(R,n,p) := \inf_{A \in \mathbb{R}^{n \times p}} \inf_{\mathcal{R}: \mathbb{R}^n \to \mathbb{R}^p} \sup_{x \in \mathcal{B}_q(R,p)} \|\mathcal{R}(Ax) - x\|_2, \tag{1.35}$$

where the second infimum is over all possible recovery algorithms that map a vector of measurements $y \in \mathbb{R}^n$ to a signal in $\mathbb{R}^p$. Remarkably, it turns out that the precise order of $\mathcal{M}_q(R,n,p)$ can be calculated from approximation theory, as given in the following theorem due to [Fou+10].

Theorem 1.5 ([Fou+10])**.** *Suppose $n \geq \log(ep/n)$, and $q \in (0,1)$. Then, the minimax ℓ_2 approximation error $\mathcal{M}_q(R,n,p)$ satisfies*

$$\mathcal{M}_q(R,n,p) \asymp R \cdot \left(\tfrac{\log(ep/n)}{n}\right)^{1/q-1/2}. \tag{1.36}$$

The result is a direct consequence of Theorem 1.1 and Proposition 1.2 of the paper [Fou+10], which builds on the early foundational work of Kashin [Kas77], Garnaev, and Gluskin [GG84].

By comparing the upper bound in line (1.34) with the minimax rate (1.36), we reach the conclusion that BPDN is rate optimal for noiseless measurements and signals in $\mathcal{B}_q(R,p)$. Further discussion of optimality properties of recovery algorithms may be found in the book [FR13].

1.5 Contributions and organization of the dissertation

The remainder of the dissertation is organized in two chapters. In Chapter 2, we study the residual bootstrap (RB) method for high-dimensional regression, and in Chapter 3, we focus on the issue of unknown sparsity in compressed sensing. The proofs of the theoretical results for each of these chapters may be found in Appendices A and B respectively. Also, we note that much of the material in Chapter 2 appeared previously in our work [Lop14] published in the proceedings of the 2014 NIPS conference, and similarly, the material in Chapter 3 builds on our work [Lop13] published in the proceedings of the 2013 ICML conference.

Chapter 2. When regression coefficients are estimated via least squares, the results of Bickel and Freedman, presented in Section 1.3, show that the RB method consistently approximates the laws of contrasts, provided that $p \ll n$, where the design matrix is of size $n \times p$. Up to now, relatively little work has considered how additional structure in the linear model may extend the validity of RB to the setting where $p/n \asymp 1$. In this setting, we propose a version of RB that resamples residuals obtained from ridge regression. Our main structural assumption on the design matrix is that it is nearly low rank — an assumption that is satisfied in many applied regression problems, and one that can be inspected directly from the observed design matrix. Under a few extra technical assumptions, we derive a simple criterion for ensuring that RB consistently approximates the laws of contrasts of the form $c^\top(\widehat{\beta}_\rho - \beta)$, where $\widehat{\beta}_\rho$ is a ridge regression estimator and β is the vector of true coefficients. We then specialize this result to study confidence intervals for mean response values $X_i^\top \beta$, where $X_i^\top$ is the ith row of the design. More precisely, we show that conditionally on a Gaussian design with near low-rank structure, RB *simultaneously* approximates all of the laws $X_i^\top(\widehat{\beta}_\rho - \beta)$, $i = 1, \dots, n$. This result is of particular theoretical significance, due to the fact that the ordinary RB method is known to fail when approximating the laws of such contrasts if least-squares residuals are used, as summarized in Section 1.3. Furthermore, the assumptions underlying our consistency results are mild to the extent that they do not depend on the existence of a limiting distribution for the contrasts $c^\top(\widehat{\beta}_\rho - \beta)$, and also do not require the vector β to be sparse.

Chapter 3. The theory of Compressed Sensing (CS) asserts that an unknown signal $x \in \mathbb{R}^p$ can be accurately recovered from an underdetermined set of linear measurements, provided that x is sufficiently sparse. However, in applications, the degree of sparsity $\|x\|_0$ is typically unknown, and the problem of directly estimating $\|x\|_0$ has been a longstanding gap between theory and practice. A closely related issue is that $\|x\|_0$ is a highly idealized measure of sparsity, and for real signals with entries not exactly equal to 0, the value $\|x\|_0 = p$ is not a useful description of compressibility. In our previous work toward addressing these problems, [Lop13], we considered an alternative measure of "soft" sparsity, $\|x\|_1^2/\|x\|_2^2$, and designed a procedure to estimate $\|x\|_1^2/\|x\|_2^2$ that does not rely on sparsity assumptions.

The present work offers a new deconvolution-based method for estimating unknown sparsity, which has wider applicability and sharper theoretical guarantees. Whereas our earlier work was limited to estimating $\|x\|_1^2/\|x\|_2^2$, this chapter introduces a family of entropy-based sparsity measures $s_q(x) := \big(\frac{\|x\|_q}{\|x\|_1}\big)^{\frac{q}{1-q}}$ parameterized by $q \in [0, \infty]$. This family interpolates between $\|x\|_0 = s_0(x)$ and $\|x\|_1^2/\|x\|_2^2 = s_2(x)$ as q ranges over $[0, 2]$, and our proposed method allows $s_q(x)$ to be estimated for all $q \in (0,2] \setminus \{1\}$. In particular, $\|x\|_0$ can be approximated via an estimate of $s_q(x)$ when q is small. Two other advantages of the new approach are that it handles measurement noise with *infinite variance*, and that it yields confidence intervals for $s_q(x)$ with asymptotically exact coverage probability.

In addition to confidence intervals, we analyze several other aspects of our proposed estimator $\widehat{s}_q(x)$. An important property of $\widehat{s}_q(x)$ is that its relative error converges at the *dimension-free* rate of $1/\sqrt{n}$. This means that using only $n = \mathcal{O}(1)$ measurements, $s_q(x)$ can be estimated to any fixed degree of relative error, even when p is arbitrarily large. Next, in connection with recovering the full signal x, we give new insight into the role of $s_2(x)$ by deriving matching upper and lower bounds on the relative error of the Basis Pursuit Denoising (BPDN) algorithm, at rate $\sqrt{s_2(x)\log(pe/n)/n}$. Finally, since our proposed method is based on randomized measurements, we show that the use of randomization is essential. Specifically, we show that the minimax relative error for estimating $s_q(x)$ with noiseless deterministic measurements is at least of order 1 when $n < p$ and $q \in [0,2]$.

Chapter 2

A Residual Bootstrap in High Dimensions

In this chapter, we focus our attention on high-dimensional linear regression, and our aim is to know when the residual bootstrap (RB) method consistently approximates the laws of *linear contrasts.*

To specify the model, suppose that we observe a response vector $Y \in \mathbb{R}^n$, generated according to

$$Y = X\beta + \varepsilon, \tag{2.1}$$

where $X \in \mathbb{R}^{n\times p}$ is a given design matrix, $\beta \in \mathbb{R}^p$ is an unknown vector of coefficients, and the error variables $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ are drawn i.i.d. from an unknown distribution F_0, with mean 0 and unknown variance $\sigma^2 < \infty$. As is conventional in high-dimensional statistics, we assume the model (2.1) is embedded in a sequence of models indexed by n. Hence, we allow X, β, and p to vary implicitly with n. We will leave p/n unconstrained until Section 2.2, where we will assume $p/n \asymp 1$ in Theorem 2.3, and then in Section 2.2, we will assume further that p/n is bounded strictly between 0 and 1.[1] The distribution F_0 is fixed with respect to n, and none of our results require F_0 to have more than four moments.

Although we are primarily interested in cases where the design matrix X is deterministic, we will also study the performance of the bootstrap conditionally on a Gaussian design. For this reason, we will use the symbol $\mathbb{E}[\dots|X]$ even when the design is non-random so that confusion does not arise in relating different sections of the chapter. Likewise, the symbol $\mathbb{E}[\dots]$ refers to unconditional expectation over all sources of randomness. Whenever the design is random, we will assume $X \perp\!\!\!\perp \varepsilon$, denoting the distribution of X by $\mathbb{P}_X$, and the distribution of ε by $\mathbb{P}_\varepsilon$.

Within the context of the regression, we will be focused on linear contrasts $c^\top(\widehat{\beta} - \beta)$, where $c \in \mathbb{R}^p$ is a fixed vector and $\widehat{\beta} \in \mathbb{R}^p$ is an estimate of β. The importance of contrasts arises from the fact that they unify many questions about a linear model. For instance, testing the significance of the ith coefficient β_i may be addressed by choosing c to be the

[1] We plan to remove this restriction in a forthcoming version of this work.

standard basis vector $c^\top = e_i^\top$. Another important problem is quantifying the uncertainty of point predictions, which may be addressed by choosing $c^\top = X_i^\top$, i.e. the ith row of the design matrix. In this case, an approximation to the law of the contrast leads to a confidence interval for the mean response value $\mathbb{E}[Y_i] = X_i^\top \beta$. Further applications of contrasts occur in the broad topic of ANOVA [LR05].

Intuition for structure and regularization in RB. The following two paragraphs explain the core conceptual aspects of the chapter. To understand the role of regularization in applying RB to high-dimensional regression, it is helpful to think of RB in terms of two ideas. First, if $\widehat{\beta}_{\text{LS}}$ denotes the ordinary least squares estimator, then it is a simple but important fact that contrasts can be written as $c^\top(\widehat{\beta}_{\text{LS}} - \beta) = a^\top \varepsilon$ where $a^\top := c^\top(X^\top X)^{-1}X^\top$. Hence, if it were possible to sample directly from F_0, then the law of any such contrast could be easily determined. Since F_0 is unknown, the second key idea is to use the residuals of *some* estimator $\widehat{\beta}$ as a proxy for samples from F_0. When $p \ll n$, the least-squares residuals are a good proxy [Fre81; BF83]. However, it is well-known that least-squares tends to overfit when $p/n \asymp 1$. When $\widehat{\beta}_{\text{LS}}$ fits "too well", this means that its residuals are "too small", and hence they give a poor proxy for F_0. Therefore, by using a regularized estimator $\widehat{\beta}$, overfitting can be avoided, and the residuals of $\widehat{\beta}$ may offer a better way of obtaining "approximate samples" from F_0.

The form of regularized regression we will focus on is *ridge regression*:

$$\widehat{\beta}_\rho := (X^\top X + \rho I_{p\times p})^{-1} X^\top Y, \tag{2.2}$$

where $\rho > 0$ is a user-specificed regularization parameter. As will be seen in Sections 2.2 and 2.2, the residuals obtained from ridge regression lead to a particularly good approximation of F_0 when the design matrix X is nearly low-rank, in the sense that most of its singular values are close to 0. In essence, this condition is a form of sparsity, since it implies that the rows of X nearly lie in a low-dimensional subspace of $\mathbb{R}^p$. However, this type of structural condition has a significant advantage over the the more well-studied assumption that β is sparse. Namely, the assumption that X is nearly low-rank can be inspected directly in practice — whereas sparsity in β is typically unverifiable without special control over the design matrix (cf. Chapter 3). In fact, our results will impose no conditions on β, other than that $\|\beta\|_2$ remains bounded as $(n, p) \to \infty$. Finally, it is worth noting that the occurrence of near low-rank design matrices is actually very common in applications, and is often referred to as *collinearity* [DS98, ch. 17].

Contributions and outline. The primary contribution of this chapter is a complement to the work of Bickel and Freedman [BF83] (hereafter B&F 1983) — who showed that in general, the RB method fails to approximate the laws of least-squares contrasts $c^\top(\widehat{\beta}_{\text{LS}} - \beta)$ when $p/n \asymp 1$. (See the discussion in Section 1.3 of Chapter 1 for additional details.) Instead, we develop an alternative set of results, proving that even when $p/n \asymp 1$, RB can successfully approximate the laws of "ridged contrasts" $c^\top(\widehat{\beta}_\rho - \beta)$ for many choices of $c \in \mathbb{R}^p$

— provided that the design matrix X is nearly low rank and that the resampled residuals are obtained from a ridge estimator. A particularly interesting consequence of our work is that RB successfully approximates the law $c^\top(\widehat{\beta}_\rho - \beta)$ for a certain choice of c that was shown in B&F 1983 to "break" RB when applied to least-squares. Specifically, such a c can be chosen as one of the rows of X with a high *leverage score*. This example corresponds to the practical problem of setting confidence intervals for mean response values $\mathbb{E}[Y_i] = X_i^\top \beta$. (Additional background is given in Section 1.3 of Chapter 1, as well as in Section 2.2 in the current chapter). Lastly, from a technical point of view, a third notable aspect of our results is that they are formulated in terms of the Mallows-ℓ_2 metric, which frees us from having to rely on the existence of a limiting distribution.

Apart from B&F 1983, the most closely related works we are aware of are the recent papers [CL13], [LY13], and [EP15], which also consider RB in the high-dimensional setting. The first two works differs from ours insofar as they focus on role of sparsity in β and do not make use of low-rank structure in the design. (Our work deals only with structure in the design and imposes no sparsity assumptions on β.) Lastly, the theoretical results in the third paper [EP15] concentrate on the failure of RB in the presence of unstructured designs.

The remainder of the chapter is organized as follows. In Section 2.1, we formulate the problem of approximating the laws of contrasts, and describe our proposed methodology for RB based on ridge regression. Then, in Section 2.2 we state several results that lay the groundwork for Theorem 2.4, which shows that that RB can successfully approximate all of the laws $\mathcal{L}(X_i^\top(\widehat{\beta}_\rho - \beta)|X)$, $i = 1, \dots, n$, conditionally on a Gaussian design. Proofs are given in Appendix A.

2.1 Problem setup and methodology

Problem setup. For any $c \in \mathbb{R}^p$, it is clear that conditionally on X, the law of $c^\top(\widehat{\beta}_\rho - \beta)$ is completely determined by F_0, and hence it makes sense to use the notation

$$\Psi_\rho(F_0; c) := \mathcal{L}\big(c^\top(\widehat{\beta}_\rho - \beta) \,\big|\, X\big). \tag{2.3}$$

The problem we aim to solve is to approximate the distribution $\Psi_\rho(F_0; c)$ for suitable choices of c.

Residual bootstrap for ridge regression. Here, we briefly explain the steps involved in the residual bootstrap procedure, applied to the ridge estimator $\widehat{\beta}_\rho$ of β. To proceed somewhat indirectly, consider the following "bias-variance" decomposition of $\Psi_\rho(F_0; c)$, conditionally on X,

$$\Psi_\rho(F_0; c) = \underbrace{\mathcal{L}\big(c^\top\big(\widehat{\beta}_\rho - \mathbb{E}[\widehat{\beta}_\rho|X]\big) \,\big|\, X\big)}_{=:\ \Phi_\rho(F_0;c)} + \underbrace{c^\top\big(\mathbb{E}[\widehat{\beta}_\rho|X] - \beta\big)}_{=:\,\text{bias}(\Phi_\rho(F_0;c))}. \tag{2.4}$$

Note that the distribution $\Phi(F_0; c)$ has mean zero, so that the second term on the right side is the bias of $\Phi_\rho(F_0; c)$ as an estimator of $\Psi_\rho(F_0; c)$. Furthermore, the distribution

$\Phi_\rho(F_0;c)$ may be viewed as the "variance component" of $\Psi_\rho(F_0;c)$. We will be interested in situations where the regularization parameter ρ may be chosen small enough so that the bias component is small. In this case, one has $\Psi_\rho(F_0;c) \approx \Phi_\rho(F_0;c)$, and then it is enough to find an approximation to the law $\Phi_\rho(F_0;c)$, which is unknown. To this end, a simple manipulation of $c^\top(\widehat{\beta}_\rho - \mathbb{E}[\widehat{\beta}_\rho|X])$ leads to

$$\Phi_\rho(F_0;c) = \mathcal{L}(c^\top(X^\top X + \rho I_{p\times p})^{-1}X^\top \varepsilon \,\big|\, X). \tag{2.5}$$

Now, to approximate $\Phi_\rho(F_0;c)$, let $\widehat{F}$ be any centered estimate of F_0. (Typically, $\widehat{F}$ is obtained by using the centered residuals of some estimator of β, but this is not necessary in general.) Also, let $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*) \in \mathbb{R}^n$ be an i.i.d. sample from $\widehat{F}$. Then, replacing ε with ε^* in line (2.5) yields

$$\Phi_\rho(\widehat{F};c) = \mathcal{L}(c^\top(X^\top X + \rho I_{p\times p})^{-1}X^\top \varepsilon^* \,\big|\, X). \tag{2.6}$$

At this point, we define the (random) measure $\Phi_\rho(\widehat{F};c)$ to be the RB approximation to $\Phi_\rho(F_0;c)$. Hence, it is clear that the RB approximation is simply a "plug-in rule".

A two-stage approach. An important feature of the procedure just described is that we are free to use any centered estimator $\widehat{F}$ of F_0. This fact offers substantial flexibility in approximating $\Psi_\rho(F_0;c)$. One way of exploiting this flexibility is to consider a two-stage approach, where a "pilot" ridge estimator $\widehat{\beta}_\varrho$ is used to first compute residuals whose centered empirical distribution function is $\widehat{F}_\varrho$, say. Then, in the second stage, the distribution $\widehat{F}_\varrho$ is used to approximate $\Phi_\rho(F_0;c)$ via the relation (2.6). (Note that such an approach involves two distinct regularization parameters ρ and ϱ.)

To be more detailed, if $(\widehat{e}_1(\varrho), \dots, \widehat{e}_n(\varrho)) = \widehat{e}(\varrho) := Y - X\widehat{\beta}_\varrho$ are the residuals of $\widehat{\beta}_\varrho$, then we define $\widehat{F}_\varrho$ to be the distribution that places mass $1/n$ at each of the values $\widehat{e}_i(\varrho) - \bar{e}(\varrho)$ with $\bar{e}(\varrho) := \frac{1}{n}\sum_{i=1}^n \widehat{e}_i(\varrho)$. Here, it is important to notice that the value ϱ is chosen to optimize $\widehat{F}_\varrho$ as an approximation to F_0. By contrast, the choice of ρ depends on the relative importance of width and coverage probability for confidence intervals based on $\Phi_\rho(\widehat{F}_\varrho;c)$. Theorems 2.1, 2.3, and 2.4 will offer some guidance in selecting ϱ and ρ.

Resampling algorithm. To summarize the discussion above, if B is user-specified number of bootstrap replicates, our proposed method for approximating $\Psi_\rho(F_0;c)$ is given below.

1. Select ρ and ϱ, and compute the residuals $\widehat{e}(\varrho) = Y - X\widehat{\beta}_\varrho$.
2. Compute the centered distribution function $\widehat{F}_\varrho$, putting mass $1/n$ at each $\widehat{e}_i(\varrho) - \bar{e}(\varrho)$.
3. For $j = 1, \dots, B$:
    - Draw a vector $\varepsilon^* \in \mathbb{R}^n$ of n i.i.d. samples from $\widehat{F}_\varrho$.

- Compute $z_j := c^\top (X^\top X + \rho I_{p\times p})^{-1} X^\top \varepsilon^*$.

4. Return the empirical distribution of $z_1, \dots, z_B$.

Clearly, as $B \to \infty$, the empirical distribution of $z_1, \dots, z_B$ converges weakly to $\Phi_\rho(\widehat{F_\varrho}; c)$, with probability 1. As is conventional, our theoretical analysis in the next section will ignore Monte Carlo issues, and address only the performance of $\Phi_\rho(\widehat{F_\varrho}; c)$ as an approximation to $\Psi_\rho(F_0; c)$.

2.2 Main results

A bias-variance decomposition for bootstrap approximation

To give some notation for analyzing the bias-variance decomposition of $\Psi_\rho(F_0; c)$ in line (2.4), we define the following quantities based upon the ridge estimator $\widehat{\beta}_\rho$. Namely, the variance is

$$v_\rho = v_\rho(X; c) := \mathrm{var}(\Psi_\rho(F_0; c)|X) = \sigma^2 \|c^\top (X^\top X + \rho I_{p\times p})^{-1} X^\top\|_2^2.$$

To express the bias of $\Phi_\rho(F_0; c)$, we define the vector $\delta(X) \in \mathbb{R}^p$ according to

$$\delta(X) := \beta - \mathbb{E}[\widehat{\beta}_\rho] = \big[I_{p\times p} - (X^\top X + \rho I_{p\times p})^{-1} X^\top X\big]\beta, \tag{2.7}$$

and then put

$$b_\rho^2 = b_\rho^2(X; c) := \mathrm{bias}^2(\Phi_\rho(F_0; c)) = (c^\top \delta(X))^2. \tag{2.8}$$

We will sometimes omit the arguments of v_ρ and b_ρ^2 to lighten notation. Note that $v_\rho(X; c)$ does not depend on β, and $b_\rho^2(X; c)$ only depends on β through $\delta(X)$.

The following result gives a regularized and high-dimensional extension of some lemmas in Freedman's early work [Fre81] on RB for least squares. The result does not restrict the size of p/n, and does not place any structural assumptions on the design matrix, or on the true parameter β. Also, we make use of the Mallow (Kantorovich) metric defined in Section 1.3 of Chapter 1.

Theorem 2.1 (consistency criterion). *Suppose $X \in \mathbb{R}^{n\times p}$ is fixed. Let $\widehat{F}$ be any estimator of F_0, and let $c \in \mathbb{R}^p$ be any vector such that $v_\rho = v_\rho(X; c) \neq 0$. Then with $\mathbb{P}_\varepsilon$-probability 1, the following inequality holds for every $n \geq 1$, and every $\rho > 0$,*

$$d_2^2\Big(\tfrac{1}{\sqrt{v_\rho}}\Psi_\rho(F_0; c), \tfrac{1}{\sqrt{v_\rho}}\Phi_\rho(\widehat{F}; c)\Big) \leq \tfrac{1}{\sigma^2} d_2^2(F_0, \widehat{F}) + \tfrac{b_\rho^2}{v_\rho}. \tag{2.9}$$

Remarks. Observe that the normalization $1/\sqrt{v_\rho}$ ensures that the bound is non-trivial, since the distribution $\Psi_\rho(F_0; c)/\sqrt{v_\rho}$ has variance equal to 1 for all n (and hence does not become degenerate for large n). To consider the choice of ρ, it is simple to verify that the ratio b_ρ^2/v_ρ decreases monotonically as ρ decreases. Note also that as ρ becomes small, the variance v_ρ becomes large, and likewise, confidence intervals based on $\Phi_\rho(\widehat{F}; c)$ become wider. In other words, there is a trade-off between the width of the confidence interval and the size of the bound (2.9).

Sufficient conditions for consistency of RB. An important practical aspect of Theorem 2.1 is that for any given contrast c, the variance $v_\rho(X; c)$ can be easily estimated, since it only requires an estimate of σ^2, which can be obtained from $\widehat{F}$. Consequently, whenever theoretical bounds on $d_2^2(F_0, \widehat{F})$ and $b_\rho^2(X; c)$ are available, the right side of line (2.9) can be controlled. In this way, Theorem 2.1 offers a simple route for guaranteeing that RB is consistent. In Sections 2.2 and 2.2 to follow, we derive a bound on $\mathbb{E}[d_2^2(F_0, \widehat{F})|X]$ in the case where $\widehat{F}$ is chosen to be $\widehat{F}_\varrho$. Later on in Section 2.2, we study RB consistency in the context of prediction with a Gaussian design, and there we derive high probability bounds on both $v_\rho(X; c)$ and $b_\rho^2(X; c)$ where c is a particular row of X.

A link between bootstrap consistency and MSPE

If $\widehat{\beta}$ is an estimator of β, its mean-squared prediction error (MSPE), conditionally on X, is defined as

$$\mathrm{mspe}(\widehat{\beta}\,|X) := \tfrac{1}{n}\mathbb{E}\big[\|X(\widehat{\beta} - \beta)\|_2^2 \;\big|\; X\big]. \tag{2.10}$$

The previous subsection showed that in-law approximation of contrasts is closely tied to the approximation of F_0. We now take a second step of showing that if the centered residuals of an estimator $\widehat{\beta}$ are used to approximate F_0, then the quality of this approximation can be bounded naturally in terms of $\mathrm{mspe}(\widehat{\beta}\,|X)$. This result applies to any estimator $\widehat{\beta}$ computed from the observations (2.1).

Theorem 2.2. *Suppose $X \in \mathbb{R}^{n\times p}$ is fixed. Let $\widehat{\beta}$ be any estimator of β, and let $\widehat{F}$ be the empirical distribution of the centered residuals of $\widehat{\beta}$. Also, let F_n denote the empirical distribution of n i.i.d. samples from F_0. Then for every $n \geq 1$,*

$$\mathbb{E}\left[d_2^2(\widehat{F}, F_0) \;\big|\; X\right] \leq 2\,\mathrm{mspe}(\widehat{\beta}\,|X) + 2\,\mathbb{E}[d_2^2(F_n, F_0)] + \tfrac{2\sigma^2}{n}. \tag{2.11}$$

Remarks. As we will see in the next section, the MSPE of ridge regression can be bounded in a sharp way when the design matrix is approximately low rank, and there we will analyze $\mathrm{mspe}(\widehat{\beta}_\varrho|X)$ for the pilot estimator. Consequently, when near low-rank structure is available, the only remaining issue in controlling the right side of line (2.11) is to bound the quantity $\mathbb{E}[d_2^2(F_n, F_0)|X]$. The recent work of Bobkov and Ledoux [BL14] provides an in-depth study of this question, and they derive a variety of bounds under different tail conditions on F_0. We summarize one of their results below.

Lemma 2.1 (Bobkov and Ledoux, 2014). *If F_0 has a finite fourth moment, then*

$$\mathbb{E}[d_2^2(F_n, F_0)] \lesssim \log(n)n^{-1/2}. \tag{2.12}$$

Remarks. The fact that the *squared* distance is bounded from above at the rate $\log(n)n^{-1/2}$ is an indication that d_2 is a rather strong metric on distributions. For a detailed discussion of this result, see Corollaries 7.17 and 7.18 in the paper [BL14]. Although it is possible to obtain faster rates when more stringent tail conditions are placed on F_0, we will only need a fourth moment, since the $\text{mspe}(\widehat{\beta}|X)$ term in Theorem 2.2 will often have a slower rate than $\log(n)n^{-1/2}$, as discussed in the next section.

Consistency of ridge regression in MSPE for near low rank designs

In this subsection, we show that when the tuning parameter ϱ is set at a suitable rate, the pilot ridge estimator $\widehat{\beta}_\varrho$ is consistent in MSPE when the design matrix is near low-rank — even when p/n is large, and without any sparsity constraints on β. We now state some assumptions, using $\widehat{\Sigma} = \frac{1}{n}X^\top X$ to denote the sample covariance matrix.

A2.1. *There is a number $\nu > 0$, and absolute constants $\kappa_1, \kappa_2 > 0$, such that*

$$\kappa_1 i^{-\nu} \leq \lambda_i(\widehat{\Sigma}) \leq \kappa_2 i^{-\nu} \qquad \textit{for all} \quad i = 1, \dots, n \wedge p.$$

A 2.2. *There are absolute constants $\theta, \gamma > 0$, such that for every $n \geq 1$, $\frac{\varrho}{n} = n^{-\theta}$ and $\frac{\varrho}{n} = n^{-\gamma}$.*

A2.3. *The vector $\beta \in \mathbb{R}^p$ satisfies $\|\beta\|_2 \lesssim 1$.*

Due to Theorem 2.2, the following bound shows that the residuals of $\widehat{\beta}_\varrho$ may be used to extract a consistent approximation to F_0. Two other notable features of the bound are that it is *non-asymptotic* and *dimension-free.*

Theorem 2.3. *Suppose that $X \in \mathbb{R}^{n\times p}$ is fixed and that assumptions **A2.1**–**A2.3** hold, with $p/n \asymp 1$. Assume further that θ is chosen as $\theta = \frac{2\nu}{3}$ when $\nu \in (0, \frac{1}{2})$, and $\theta = \frac{\nu}{\nu+1}$ when $\nu > \frac{1}{2}$. Then,*

$$\text{mspe}(\widehat{\beta}_\varrho|X) \lesssim \begin{cases} n^{-\frac{2\nu}{3}} & \textit{if} \quad \nu \in (0, \frac{1}{2}), \\ n^{-\frac{\nu}{\nu+1}} & \textit{if} \quad \nu > \frac{1}{2}. \end{cases} \tag{2.13}$$

Also, both bounds in (2.13) *are tight in the sense that β can be chosen so that $\widehat{\beta}_\varrho$ attains either rate.*

Remarks. Since the eigenvalues $\lambda_i(\widehat{\Sigma})$ are observable, they may be used in principle to estimate ν and guide the selection of $\varrho/n = n^{-\theta}$. However, from a practical point of view, our experience with simulations suggests it is easier to select ϱ via cross-validation in numerical experiments, rather than via an estimate of ν.

A link with Pinsker's Theorem. In the particular case when F_0 is a centered Gaussian distribution, the "prediction problem" of estimating $X\beta$ is very similar to estimating the mean parameters of a Gaussian sequence model, with error measured in the ℓ_2 norm. In the alternative sequence-model format, the decay condition on the eigenvalues of $\frac{1}{n}X^\top X$ translates into an ellipsoid constraint on the mean parameter sequence [Tsy09; Was06]. For this reason, Theorem 2.3 may be viewed as "regression version" of ℓ_2 error bounds for the sequence model under an ellipsoid constraint (cf. Pinsker's Theorem, [Tsy09; Was06]). Due to the fact that the latter problem has a very well developed literature, there may be various "neighboring results" elsewhere. Nevertheless, we could not find a direct reference for our stated MSPE bound in the current setup. For the purposes of our work in this paper, the more important point to take away from Theorem 2.3 is that it can be coupled with Theorem 2.2 for proving consistency of RB.

Confidence intervals for mean responses, conditionally on a Gaussian design

In this section, we consider the situation where the design matrix X has rows $X_i^\top \in \mathbb{R}^p$ drawn i.i.d. from a multivariate normal distribution $N(0, \Sigma)$, with $X \perp\!\!\!\perp \varepsilon$. (The covariance matrix Σ may vary with n.) Conditionally on a realization of X, we analyze the RB approximation of the laws $\Psi_\rho(F_0; X_i) = \mathcal{L}(X_i^\top(\widehat{\beta}_\rho - \beta)|X)$. As discussed in Section 3.1, this corresponds to the problem of setting confidence intervals for the mean responses $\mathbb{E}[Y_i] = X_i^\top\beta$. Assuming that the population eigenvalues $\lambda_i(\Sigma)$ obey a decay condition, we show below in Theorem 2.4 that RB succeeds with high $\mathbb{P}_X$-probability. Moreover, this consistency statement holds for all of the laws $\Psi_\rho(F_0; X_i)$ *simultaneously.* That is, among the n distinct laws $\Psi_\rho(F_0; X_i)$, $i = 1, \dots, n$, even the worst bootstrap approximation is still consistent.

In addition to the applicability of setting confidence intervals for mean response values, our interest in approximating the laws $\Psi_\rho(F_0; X_i)$ arises from the fact that the ordinary RB based on least squares is known to fail in general for contrasts of this type. As discussed in Section 1.3 of Chapter 1, the specific choice of the vector $X_i^\top$ that breaks the ordinary RB depends on the behavior of the diagonal entries of the hat matrix $H := X(X^\top X)^{-1}X^\top$, when $X^\top X$ is invertible. Let λ_n denote the (discrete) distribution that places mass $1/n$ at each of the values $1 - H_{ii}$ for $i = 1, \dots, n$. When X is a Gaussian design, and $p/n \to \kappa \in (0, 1)$ as $(n, p) \to \infty$, it follows from Proposition A.1 in Appendix A.5 that with probability 1, the distributions λ_n converge weakly to a point mass at the value $1 - \kappa$. In this situation, Theorem 1.2 in Chapter 1 shows that for any subsequence $\{i_n\}$ along which $H_{i_n i_n} \to \kappa$ (and at least one must exist), the contrast $c^\top = X_{i_n}^\top$ leads to failure of the ordinary RB (conditionally on the design). Interestingly, it turns out that for Gaussian designs, $\max_{1 \le i \le n} |H_{ii} - p/n| \to 0$ almost surely, and so in fact, choosing $c^\top$ to be any row of the design leads to the failure of the ordinary RB.

We now state some population-level assumptions in preparation for the results of this section.

A2.4. *The operator norm of* $\Sigma \in \mathbb{R}^{p\times p}$ *satisfies* $\|\Sigma\|_{\text{op}} \lesssim 1$.

Next, we impose a decay condition on the eigenvalues of Σ. This condition also ensures that Σ is invertible for each fixed p — even though the bottom eigenvalue may become arbitrarily small as p becomes large. It is also important to notice that we now use η for the decay exponent of the population eigenvalues, whereas we used ν when describing the sample eigenvalues in the previous section.

A2.5. *There is a number* $\eta > 0$*, and absolute constants* $k_1, k_2 > 0$*, such that for all* $i = 1, \dots, p$*,*

$$k_1 i^{-\eta} \leq \lambda_i(\Sigma) \leq k_2 i^{-\eta}.$$

A2.6. *There are absolute constants* $k_3, k_4 \in (0,1)$ *such that for all* $n \geq 3$*, we have the bounds* $k_3 \leq \frac{p}{n} \leq k_4$ *and* $p \leq n-2$.

The following lemma collects most of the effort needed in proving our final result in Theorem 2.4. Here it is also helpful to recall the notation $\rho/n = n^{-\gamma}$ and $\varrho/n = n^{-\theta}$ from Assumption 2.2.

Lemma 2.2. *Suppose that the matrix* $X \in \mathbb{R}^{n\times p}$ *has rows* $X_i^\top$ *drawn i.i.d. from* $N(0, \Sigma)$*, and that assumptions* ***A2.2–A2.6*** *hold. Furthermore, assume that* γ *chosen so that* $0 < \gamma < \min\{\eta, 1\}$*. Then, the statements below are true.*
(i) (bias inequality)
Fix any $\tau > 0$*. Then, there is an absolute constant* $\kappa_0 > 0$*, such that for all large* n*, the following event holds with* $\mathbb{P}_X$*-probability at least* $1 - n^{-\tau} - ne^{-n/16}$*,*

$$\max_{1\leq i\leq n} b_\rho^2(X; X_i) \ \leq \kappa_0 \cdot n^{-\gamma} \cdot (\tau+1)\log(n+2). \tag{2.14}$$

(ii) (variance inequality)
There are absolute constants $\kappa_1, \kappa_2 > 0$ *such that for all large* n*, the following event holds with* $\mathbb{P}_X$*-probability at least* $1 - 4n\exp(-\kappa_1 n^{\frac{\gamma}{\eta}})$*,*

$$\max_{1\leq i\leq n} \tfrac{1}{v_\rho(X;X_i)} \leq \kappa_2 n^{1-\frac{\gamma}{\eta}}. \tag{2.15}$$

(iii) (mspe inequalities)
Suppose that θ *is chosen as* $\theta = 2\eta/3$ *when* $\eta \in (0, \frac{1}{2})$*, and that* θ *is chosen as* $\theta = \frac{\eta}{1+\eta}$ *when* $\eta > \frac{1}{2}$*. Then, there are absolute constants* $\kappa_3, \kappa_4, \kappa_5, \kappa_6 > 0$ *such that for all large* n*,*

$$\text{mspe}(\widehat{\beta}_\varrho|X) \leq \begin{cases} \kappa_4 n^{-\frac{2\eta}{3}} \ \textit{with } \mathbb{P}_X\textit{-prob. at least } \ 1-\exp(-\kappa_3 n^{2-4\eta/3}), & \textit{if} \ \ \eta \in (0, \frac{1}{2}) \\ \kappa_6 n^{-\frac{\eta}{\eta+1}} \ \textit{with } \mathbb{P}_X\textit{-prob. at least } \ 1-\exp(-\kappa_5 n^{\frac{2}{1+\eta}}), & \textit{if} \ \ \eta > \frac{1}{2}. \end{cases}$$

Remarks. Note that the two rates in part (iii) coincide as η approaches 1/2. At a conceptual level, the entire lemma may be explained in relatively simple terms. Viewing the quantities $\text{mspe}(\widehat{\beta}_\varrho|X)$, $b_\rho^2(X;X_i)$ and $v_\rho(X;X_i)$ as functionals of a Gaussian matrix, the proof involves deriving concentration bounds for each of them. Indeed, this is plausible given that these quantities are smooth functionals of X. However, the difficulty of the proof arises from the fact that they are also highly non-linear functionals of X. We now combine Lemmas 2.1 and 2.2 with Theorems 2.1 and 2.2 to show that all of the laws $\Psi_\rho(F_0;X_i)$ can be simultaneously approximated via our two-stage RB method.

Theorem 2.4. *Suppose that F_0 has a finite fourth moment, assumptions **A2.2–A2.6** hold, and γ is chosen so that $\frac{\eta}{1+\eta} < \gamma < \min\{\eta, 1\}$. Also suppose that θ is chosen as $\theta = 2\eta/3$ when $\eta \in (0, \frac{1}{2})$, and $\theta = \frac{\eta}{\eta+1}$ when $\eta > \frac{1}{2}$. Then, there is a sequence of positive numbers δ_n with $\lim_{n\to\infty} \delta_n = 0$, such that the event*

$$\mathbb{E}\Big[\max_{1\leq i\leq n} d_2^2\Big(\tfrac{1}{\sqrt{v_\rho}}\Psi_\rho(F_0;X_i),\, \tfrac{1}{\sqrt{v_\rho}}\Phi_\rho(\widehat{F}_\varrho;X_i)\Big)\,\Big|\, X\Big] \leq \delta_n \tag{2.16}$$

has $\mathbb{P}_X$-probability tending to 1 as $n \to \infty$.

Remark. Lemma 2.2 gives explicit bounds on the numbers δ_n, as well as the probabilities of the corresponding events, but we have stated the result in this way for the sake of readability.

2.3 Simulations

In four different settings of n, p, and the decay parameter η, we compared the nominal 90% confidence intervals (CIs) of four methods: "oracle", "ridge", "normal", and "OLS", to be described below. In each setting, we generated $N_1 := 100$ random designs X with i.i.d. rows drawn from $N(0,\Sigma)$, where $\lambda_j(\Sigma) = j^{-\eta}$, $j = 1,\ldots,p$, and the eigenvectors of Σ were drawn randomly by setting them to be the Q factor in a QR decomposition of a standard $p \times p$ Gaussian matrix. Then, for each realization of X, we generated $N_2 := 1000$ realizations of Y according to the model (2.1), where $\beta = \mathbf{1}/\|\mathbf{1}\|_2 \in \mathbb{R}^p$, and F_0 is the centered t distribution on 5 degrees of freedom, rescaled to have standard deviation $\sigma = 0.1$. For each X, and each corresponding Y, we considered the problem of setting a 90% CI for the mean response value $X_{i^\star}^\top \beta$, where $X_{i^\star}^\top$ is the row with the highest leverage score, i.e. $i^\star = \text{argmax}_{1\leq i\leq n} H_{ii}$ with $H = X(X^\top X)^{-1}X^\top$. This choice is motivated by the fact that it is known to break the standard RB method based on least-squares fails when $p/n \asymp 1$.[2] Below, we refer to this method as "OLS".

To describe the other three methods, "ridge" refers to the interval $[X_{i^\star}^\top\widehat{\beta}_\rho - \widehat{q}_{0.95}, X_{i^\star}^\top\widehat{\beta}_\rho - \widehat{q}_{0.05}]$, where $\widehat{q}_\alpha$ is the α% quantile of the numbers $z_1,\ldots,z_B$ computed in the proposed algorithm in Section 2.1, with $B = 1000$ and $c^\top = X_{i^\star}^\top$. To choose the parameters ρ and ϱ for a given X and Y, we first computed $\widehat{r}$ as the value that optimized the MSPE error of a

[2] See the discussion at the beginning of the previous section, as well as Section 1.3 of Chapter 1).

ridge estimator $\widehat{\beta}_r$ with respect to 5-fold cross validation; i.e. cross validation was performed for every distinct pair (X, Y). We then put $\varrho = 5\widehat{r}$ and $\rho = 0.1\widehat{r}$, as we found the prefactors 5 and 0.1 to work adequately across various settings. (Optimizing ϱ with respect to MSPE is motivated by Theorems 2.1, 2.2, and 2.3. Also, choosing ρ to be somewhat smaller than ϱ conforms with the constraints on θ and γ in Theorem 2.4.) The method "normal" refers to the CI based on the (heuristic) normal approximation $\mathcal{L}(X_{i^\star}^\top(\widehat{\beta}_\rho - \beta)|X) \approx N(0, \widehat{\tau}^2)$, where $\widehat{\tau}^2 = \widehat{\sigma}^2\|X_{i^\star}^\top(X^\top X + \rho I_{p\times p})^{-1}X^\top\|_2^2$, $\rho = 0.1\widehat{r}$, and $\widehat{\sigma}^2$ is the usual unbiased estimate of σ^2 based on OLS residuals. The "oracle" method refers to the interval $[X_{i^\star}^\top\widehat{\beta}_\rho - \tilde{q}_{0.95}, X_{i^\star}^\top\widehat{\beta}_\rho - \tilde{q}_{0.05}]$, with $\rho = 0.1\widehat{r}$, and $\tilde{q}_\alpha$ being the empirical α% quantile of $X_i^\top(\widehat{\beta}_\rho - \beta)$ over all 1000 realizations of Y based on a given X. (This accounts for the randomness in $\rho = 0.1\widehat{r}$.)

Within a given setting of the triplet (n, p, η), we refer to the "coverage" of a method as the fraction of the $N_1 \times N_2 = 10^5$ instances where the method's CI contained the parameter $X_{i^\star}^\top\beta$. Also, we refer to "width" as the average width of a method's intervals over all of the 10^5 instances. The four settings of (n, p, η) correspond to moderate/high dimension and moderate/fast decay of the eigenvalues $\lambda_i(\Sigma)$. Even in the moderate case of $p/n = 0.45$, the results show that the OLS intervals are too narrow and have coverage noticeably less than 90%. As expected, this effect becomes more pronounced when $p/n = 0.95$. The ridge and normal intervals perform reasonably well across settings, with both performing much better than OLS. However, it should be emphasized that our study of RB is motivated by the desire to gain insight into the behavior of the bootstrap in high dimensions — rather than trying to outperform particular methods. In future work, we plan to investigate the relative merits of the ridge and normal intervals in greater detail.

Table 2.1: Comparison of nominal 90% confidence intervals

		oracle	ridge	normal	OLS
setting 1	width	0.21	0.20	0.23	0.16
$n = 100,\ p = 45,\ \ \eta = 0.5$	coverage	0.90	0.87	0.91	0.81
setting 2	width	0.22	0.26	0.26	0.06
$n = 100,\ p = 95,\ \ \eta = 0.5$	coverage	0.90	0.88	0.88	0.42
setting 3	width	0.20	0.21	0.22	0.16
$n = 100,\ p = 45,\ \ \eta = 1$	coverage	0.90	0.90	0.91	0.81
setting 4	width	0.21	0.26	0.23	0.06
$n = 100,\ p = 95,\ \ \eta = 1$	coverage	0.90	0.92	0.87	0.42

Chapter 3

Unknown Sparsity in Compressed Sensing

3.1 Introduction

In this chapter, we consider the standard compressed sensing (CS) model, involving n linear measurements $y = (y_1, \dots, y_n)$, generated according to

$$y = Ax + \sigma\epsilon, \tag{3.1}$$

where $x \in \mathbb{R}^p$ is an unknown signal, $A \in \mathbb{R}^{n\times p}$ is a measurement matrix specified by the user, $\sigma\epsilon \in \mathbb{R}^n$ is a random noise vector, and $n \ll p$. The central problem of CS is to recover the signal x using only the observations y and the matrix A. Over the course of the past decade, a large body of research has shown that this seemingly ill-posed problem can be solved reliably when x is sparse. Specifically, when the sparsity level of x is measured in terms of the ℓ_0 norm $\|x\|_0 := \text{card}\{j : x_j \neq 0\}$, it is well known that if $n \gtrsim \|x\|_0 \log(p)$, then accurate recovery can be achieved with high probability when A is drawn from a suitable ensemble [Don06; CRT06; EK12; FR13]. In this way, the parameter $\|x\|_0$ is often treated as being known in much of the theoretical CS literature — despite the fact that $\|x\|_0$ is usually *unknown* in practice. Due to the fact that the sparsity parameter plays a fundamental role in CS, the issue of unknown sparsity has become recognized as gap between theory and practice [War09; Eld09; MSW08; BDB07]. Likewise, our overall focus in this chapter is the problem of estimating the unknown sparsity level of x without relying on any sparsity assumptions.

Motivations and the role of sparsity

Given that many well-developed methods are available for estimating the full signal x, or its support set $S := \{j \in \{1, \dots, p\} : x_j \neq 0\}$, it might seem surprising that the problem of estimating $\|x\|_0$ has remained largely unsettled. Indeed, given an estimate of x or S, it might seem natural to estimate $\|x\|_0$ via a “plug-in rule”, such as $\|\widehat{x}\|_0$ or $\text{card}(\widehat{S})$. However, it is important to recognize that methods for computing $\widehat{x}$ and $\widehat{S}$ generally rely on prior knowledge of $\|x\|_0$. Consequently, when using a plug-in rule to estimate $\|x\|_0$, there is a danger of circular reasoning, and as a result, the problem of estimating $\|x\|_0$ does not simply reduce to estimating x or S.

To give a more concrete sense for the importance of estimating unknown sparsity, the following list illustrates many aspects of CS where sparsity assumptions play an important role, and where it would be valuable to have an "assumption-free" estimate of $\|x\|_0$.

1. **Modeling assumptions and choice of basis.** In some signal processing applications, sparsity-based methods are not the only viable approach, and it is of basic interest to know whether not a sparse representation is justified by data. For instance, this issue has been actively studied in the areas of face recognition and image classification: [Shi+11; DHG13; Wan+14; RBL11]. In this context, an assumption-free estimate of $\|x\|_0$ would serve as a natural diagnostic tool in model development.

   A second issue that is related to model development is the choice of basis used to represent a signal. Although there are many application-specific bases (e.g. various types of wavelet bases) that often lead to sparse representations, the ability to "validate" the choice of basis has clear practical value. In this direction, an estimator of $\|x\|_0$ could be of use in comparing the relative merits of different bases.

2. **The number of measurements.** If the choice of n is too small compared to the "critical" number $n^* \approx \|x\|_0 \log(p)$, then there are known information-theoretic barriers to the accurate reconstruction of x [RWY11]. At the same time, if n is chosen to be much larger than n^*, then the measurement process is wasteful (since there are known algorithms that can reliably recover x with approximately n^* measurements [EK12]). For this reason, it not only important to ensure that $n \geq n^*$, but to choose n close to n^*.

   To deal with the selection of n, a sparsity estimate $\widehat{\|x\|}_0$ may be used in two different ways, depending on whether measurements are collected sequentially, or in a single batch. In the sequential case, an estimate of $\|x\|_0$ can be computed from a small set of "preliminary" measurements, and then the estimated value $\widehat{\|x\|}_0$ determines how many additional measurements should be collected to recover the full signal. Also, it may not even be necessary to take additional measurements, since the preliminary set may be re-used to compute $\widehat{x}$. Alternatively, if all of the measurements must be taken in one batch, the value $\widehat{\|x\|}_0$ can be used to certify whether or not enough measurements were actually taken.

3. **The measurement matrix.** The performance of recovery procedures depends heavily on the sensing matrix A. In particular, the properties of A that lead to good recovery are often directly linked to the sparsity level of x. Two specific properties that have been intensively studied are the *restricted isometry property of order k* (RIP-k), [CT05], and the *null-space property of order k* (NSP-k),[CDD09; DH01], where k is a presumed

upper bound on the sparsity level of the true signal. (See Chapter 1 for definitions and background.) Because recovery guarantees are closely tied to RIP-k and NSP-k, a growing body of work has been devoted to certifying whether or not a given matrix satisfies these properties [dE11; JN11; TN11]. When k is treated as given, this problem is already computationally difficult. Yet, when the sparsity of x is unknown, we must also remember that such a "certificate" is less meaningful if we cannot check that the value of k is in agreement with the true signal.

4. **Recovery algorithms.** When recovery algorithms are implemented, the sparsity level of x is often treated as a tuning parameter. For example, if k is a conjectured bound on $\|x\|_0$, then the Orthogonal Matching Pursuit algorithm (OMP) is typically initialized to run for k iterations [TG07]. A second example is the Lasso algorithm, which computes a solution $\widehat{x} \in \operatorname{argmin}\{\|y - Av\|_2^2 + \lambda\|v\|_1 : v \in \mathbb{R}^p\}$, for some choice of $\lambda \geq 0$. The sparsity of $\widehat{x}$ is determined by the size of λ, and in order to select the appropriate value, a family of solutions is examined over a range of λ values [TT11]. In the case of either OMP or Lasso, a sparsity estimate $\widehat{\|x\|_0}$ would reduce computation by restricting the possible choices of λ or k, and it would also ensure that the sparsity level of the solution conforms to the true signal. With particular regard to the Lasso, an indirect consequence of our sparsity estimation method (introduced in Section 3.3) is that it allows for regularization parameter to be adaptively selected when the Lasso problem is written in "primal form": $\widehat{x} \in \operatorname{argmin}\{\|y - Av\|_2^2 : v \in \mathbb{R}^p \text{ and } \|v\|_1 \leq t\}$. See Section 3.5 for further details.

A numerically stable measure of sparsity

Despite the important theoretical role of the parameter $\|x\|_0$, it has a severe practical drawback of being sensitive to small entries of x. In particular, for real signals $x \in \mathbb{R}^p$ whose entries are not exactly equal to 0, the value $\|x\|_0 = p$ is not a useful description of compressibility.

In order to estimate sparsity in a way that accounts for the instability of $\|x\|_0$, it is desirable to replace the ℓ_0 norm with a "soft" version. More precisely, we would like to identify a function of x that can be interpreted as counting the "effective number of coordinates of x", but remains stable under small perturbations. In the next subsection, we derive such a function by showing that $\|x\|_0$ is a limiting case of a more general sparsity measure based on entropy.

A link between $\|x\|_0$ and entropy

Any vector $x \in \mathbb{R}^p \setminus \{0\}$ induces a distribution $\pi(x) \in \mathbb{R}^p$ on the set of indices $\{1, \dots, p\}$, assigning mass $\pi_j(x) := |x_j|/\|x\|_1$ at index j.[1] Under this correspondence, if x places most of

[1] It is also possible to normalize $\pi(x)$ in other ways, e.g. $\pi_j(x) = |x_j|^2/\|x\|_2^2$. See the end of Section 3.1 for additional comments.

its mass at a small number of coordinates, and $J \sim \pi(x)$ is a random variable in $\{1, \dots, p\}$, then J is likely to occupy a small set of *effective states*. This means that if x is sparse, then $\pi(x)$ has low entropy. From the viewpoint of information theory, it is well known that the entropy of a distribution can be interpreted as the logarithm of the distribution's effective number of states. Likewise, it is natural to count effective coordinates of x by counting effective states of $\pi(x)$ via entropy. To this end, we define the *numerical sparsity*[2]

$$s_q(x) := \begin{cases} \exp(H_q(\pi(x))) & \text{if } \; x \neq 0 \\ 0 & \text{if } \; x = 0, \end{cases} \tag{3.2}$$

where H_q is the Rényi entropy of order $q \in [0, \infty]$. When $q \notin \{0, 1, \infty\}$, the Rényi entropy is given explicitly by

$$H_q(\pi(x)) := \tfrac{1}{1-q} \log \left(\textstyle\sum_{i=1}^p \pi_i(x)^q \right), \tag{3.3}$$

and cases of $q \in \{0, 1, \infty\}$ are defined by evaluating limits, with H_1 being the ordinary Shannon entropy. Combining the last two lines with the definition of $\pi(x)$, we see that for $x \neq 0$ and $q \notin \{0, 1, \infty\}$, the numerical sparsity may be written conveniently in terms of ℓ_q norms as

$$s_q(x) = \left(\frac{\|x\|_q}{\|x\|_1} \right)^{\frac{q}{1-q}}. \tag{3.4}$$

As with H_q, the the cases of $q \in \{0, 1, \infty\}$ are evaluated as limits:

$$s_0(x) = \lim_{q \to 0} s_q(x) \; = \; \|x\|_0 \tag{3.5}$$

$$s_1(x) = \lim_{q \to 1} s_q(x) \; = \; \exp(H_1(\pi(x))) \tag{3.6}$$

$$s_\infty(x) = \lim_{q \to \infty} s_q(x) = \; \tfrac{\|x\|_1}{\|x\|_\infty}. \tag{3.7}$$

Background on the definition of $s_q(x)$

To the best of our knowledge, the definition of numerical sparsity (3.2) in terms of Rényi entropy is new in the context of CS. However, numerous special cases and related definitions have been considered elsewhere. For instance, in the early study of wavelet bases, Coifman and Wickerhauser proposed $\exp\big(- \sum_{i=1}^p \frac{|x_i|^2}{\|x\|_2^2} \log(\frac{|x_i|^2}{\|x\|_2^2})\big)$ as a measure of effective dimension [CW92]. (See also the papers [RK99] [Don94] [HR09].) The basic difference between this quantity and $s_q(x)$ is that the Rényi entropy leads instead to a convenient ratio of norms, which will play an important role in our procedure for estimating $s_q(x)$.

Interestingly, in recent years, there has been growing interest in ratios of norms as measures of sparsity, but such ratios have generally been introduced in an ad-hoc manner, and there has not been a principled way to explain where they "come from". To this extent, our

[2] Our terminology derives from the notion of *numerical rank* coined by [RV07].

definition of $s_q(x)$ offers a way of conceptually unifying these ratios.[3] Examples of previously studied instances include $\|x\|_1^2/\|x\|_2^2$ corresponding to $q = 2$ [Lop13],[TN11], [Hoy04], $\|x\|_1/\|x\|_\infty$ corresponding to $q = \infty$ [PGC12] [DH14], as well as $(\|x\|_a/\|x\|_b)^{ab/(b-a)}$ with $a, b > 0$, which is implicitly defined in the paper [BKS14].[4] To see how the latter quantity fits in the scope of our entropy-based definition, one may consider a different normalization of the probability vector $\pi(x)$ discussed earlier. That is, if we put $\pi_j(x) = |x_j|^t/\|x\|_t^t$ for some $t > 0$, then it follows that $\exp(H_q(\pi(x))) = (\|x\|_{tq}/\|x\|_t)^{tq/(1-q)}$. Furthermore, if one chooses $t = b$ and $q = a/b$, then the two quantities match.

Outside the context of CS, the use of Rényi entropy to count the effective number of states of a distribution has been well-established in the ecology literature for a long time. There, Rényi entropy is used to count the effective number of species in a community of organisms. More specifically, if a distribution π on $\{1, \dots, p\}$ measures the relative abundance of p species in a community, then the number $\exp(H_q(\pi))$ is a standard measure of the *effective number of species* in the community. In the ecology literature, this number is known as the *Hill index* or *diversity number* of the community. We refer the reader to the papers [Hil73] and [Jos06], as well as the references therein for further details. In essence, the main conceptual ingredient needed to connect these ideas with the notion of sparsity in CS is to interpret the signal $x \in \mathbb{R}^p$ as a distribution on the set of indices $\{1, \dots, p\}$.

Properties of $s_q(x)$

The following list summarizes some of the most important properties of $s_q(x)$, and clarifies the interpretation of $s_q(x)$ as a measure of sparsity.

(i) **(continuity).** Unlike the ℓ_0 norm, the function $s_q(\cdot)$ is continuous on $\mathbb{R}^p \setminus \{0\}$ for all $q > 0$, and is hence stable under small perturbations of x.

(ii) **(range equal to $[0, p]$).** For all $x \in \mathbb{R}^p$ and all $q \in [0, \infty]$, the numerical sparsity satisfies

$$0 \leq s_q(x) \leq p.$$

This property follows from the fact that for any q, and any distribution π on $\{1, \dots, p\}$, the Rényi entropy satisfies $1 \leq H_q(\pi) \leq \log(p)$.

(iii) **(scale-invariance).** The property that $\|cx\|_0 = \|x\|_0$ for all scalars $c \neq 0$ is familiar for the ℓ_0 norm, and this generalizes to $s_q(x)$ for all $q \in [0, \infty]$. Scale-invariance encodes the idea that sparsity should be based on relative (rather than absolute) magnitudes of the entries of x.

(iv) **(lower bound on $\|x\|_0$ and monotonicity in q).** For any $x \in \mathbb{R}^p$, the function $q \mapsto s_q(x)$ is monotone decreasing on $[0, \infty]$, and interpolates between the extreme

[3] See also our discussion of analogues of $s_q(x)$ for matrix rank in Section 3.1.

[4] This quantity is implicitly suggested in the paper [BKS14] by considering a binary vector x with $\|x\|_0 = k \geq 1$, and then choosing an exponent c so that $(\|x\|_a/\|x\|_b)^c = k$.

values of $s_\infty(x)$ and $s_0(x)$. That is, for any $q' \geq q \geq 0$, we have the bounds

$$\tfrac{\|x\|_1}{\|x\|_\infty} = s_\infty(x) \leq s_{q'}(x) \leq s_q(x) \leq s_0(x) = \|x\|_0. \tag{3.8}$$

In particular, we have the general lower bound

$$s_q(x) \leq \|x\|_0. \tag{3.9}$$

The monotonicity is a direct consequence of the fact that the Rény entropy H_q is decreasing in q.

(v) **(Schur concavity).** The notion of *majorization* formalizes the idea that the coordinates of a vector $x \in \mathbb{R}^p$ are more "spread out" than those of another vector $\tilde{x} \in \mathbb{R}^p$. (See the book [MOA10] for an in-depth treatment of majorization.) If x is majorized by $\tilde{x}$, we write $x \prec \tilde{x}$, where larger vectors in this partial order have coordinates that are less spread out. From this interpretation, one might expect that if $|x| \prec |\tilde{x}|$, then $\tilde{x}$ should be sparser than x, where $|x| := (|x_1|, \ldots, |x_p|)$. It turns out that this intuition is respected by $s_q(\cdot)$, in the sense that for any $q \in [0, \infty]$,

$$|x| \prec |\tilde{x}| \implies s_q(x) \geq s_q(\tilde{x}). \tag{3.10}$$

In general, if a function f satisfies $f(x) \geq f(\tilde{x})$ for all $x \prec \tilde{x}$ with $x, \tilde{x}$ lying in a set S, then f is said to be *Schur concave* on S. Consequently, line (3.10) implies that $s_q(\cdot)$ is Schur concave on the orthant $\mathbb{R}^p_+ \setminus \{0\}$. This property follows easily from the fact that the Rényi entropy is Schur concave on the p-dimensional probability simplex.

(vi) **(Equivalence with power-law decay).** An alternative approach to measuring "soft sparsity" is through the notion of a decay constraint on the sorted coordinate magnitudes $|x|_{[1]} \geq \cdots \geq |x|_{[p]}$. Popular measures for coordinate-wise decay include *weak ℓ_q norm* constraints, and the *power-law* decay profile

$$|x|_{[i]} \propto i^{-\tau(x)}, \tag{3.11}$$

where $\tau(x) > 0$ is a decay parameter, cf. [Joh13; Can06]. Although the parameter $\tau(x)$ does not have the convenient interpretation of an effective number of coordinates, this type of sparsity model is still quite useful for describing signals with many small (but non-zero) coordinates. Interestingly, the parameters $\tau(x)$ and $s_2(x)$ are equivalent in the sense that they are linked by an explicit bijection. Specifically, there is an invertible function $\psi : (0, p) \to (0, \infty)$ such that for all $x \in \mathbb{R}^p$ satisfying the power-law condition (3.11), we have

$$\tau(x) = \psi(s_2(x)), \tag{3.12}$$

and the inverse of ψ is given by $\psi^{-1}(\tau) = \left(\sum_{i=1}^p i^{-\tau}\right)^2 \big/ \sum_{i=1}^p i^{-2\tau}$. Consequently, $\tau(x)$ is estimable whenever $s_2(x)$ is. The proof of the invertibility of ψ is given in Section B.1 of Appendix B.

The choice of q and normalization. The parameter q controls how much weight $s_q(x)$ assigns to small coordinates. When $q = 0$, an arbitrarily small coordinate is still counted as being "effective". By contrast, when $q = \infty$, a coordinate is not counted as being effective unless its magnitude is close to $\|x\|_\infty$. The choice of q is also relevant to other considerations. For instance, we will show in Section 3.2 that the case of $q = 2$ is important because signal recovery guarantees can be derived in terms of

$$s_2(x) = \frac{\|x\|_1^2}{\|x\|_2^2}. \tag{3.13}$$

In addition, the choice of q can affect the type of measurements used to estimate $s_q(x)$. In this respect, the case of $q = 2$ turns out to be attractive because our proposed method for estimating $s_2(x)$ relies on Gaussian measurements — which can be naturally *re-used* for recovering the full signal x. Furthermore, in some applications, the measurements associated with one value of q may be easier to acquire (or process) than another. In our proposed method, smaller values of q lead to measurement vectors sampled from a distribution with heavier tails. Because a vector with heavy tailed i.i.d. entries will tend to have just a few very large entries, such vectors are approximately sparse. In this way, the choice of q may enter into the design of measurement systems because it is known that sparse measurement vectors can simplify certain recovery procedures [GI10].

Apart from the choice of q, there is a second degree of freedom associated with $s_q(x)$. In defining the probability vector $\pi_j(x) = |x_j|/\|x\|_1$ earlier, we were not forced to normalize the mass of the coordinates using $\|x\|_1$, and many other normalizations are possible. Also, some normalizations may be computationally advantageous for the problem of minimizing $s_q(x)$, as discussed below.

Minimization of $s_q(x)$? Although our work in this chapter does not deal with the problem of minimizing $s_q(x)$, some readers may still naturally be curious about what can be done in this direction. It turns out that for certain values of q, or certain normalizations of the probability vector $\pi(x)$, the minimization of $s_q(x)$ may be algorithmically tractable (under suitable constraints). The recent paper [Rep+15] discusses methods for minimizing $s_2(x)$. Another example is the minimization of $s_\infty(x)$, which can be reduced to a sequence of linear programming problems [PGC12] [DH14]. Lastly, a third example deals with the normalization $\pi_j(x) = |x_j|^t/\|x\|_t^t$ with $t = 1/2$, which leads to $\exp(H_2(\pi(x)) = (\|x\|_2/\|x\|_4)^4$. In the paper [BKS14], the problem of minimizing $\|x\|_2/\|x\|_4$ has been shown to have interesting connections with sums-of-squares (SOS) optimization problems — for which efficient algorithms are available [Las09].

Graphical interpretations

The fact that $s_q(x)$ is a sensible measure of sparsity for non-idealized signals is illustrated in Figure 1 for the case of $q = 2$. In essence, if x has k large coordinates and $p - k$ small coordinates, then $s_q(x) \approx k$, whereas $\|x\|_0 = p$. In the left panel, the sorted coordinates of three different vectors in $\mathbb{R}^{100}$ are plotted. The value of $s_2(x)$ for each vector is marked with a triangle on the x-axis, which shows that $s_2(x)$ adapts well to the decay profile. This idea can be seen in a more geometric way in the right panel, which plots the the sub-level sets $\mathcal{S}_c := \{x \in \mathbb{R}^p : s_2(x) \leq c\}$ with $c = 1.1$ and $c = 1.9$ where $p = 2$. When $c \approx 1$, the vectors in $\mathcal{S}_c$ are closely aligned with the coordinate axes, and hence contain one effective coordinate. As $c \uparrow p$, the set $\mathcal{S}_c$ expands to include less sparse vectors until $\mathcal{S}_p = \mathbb{R}^p$.

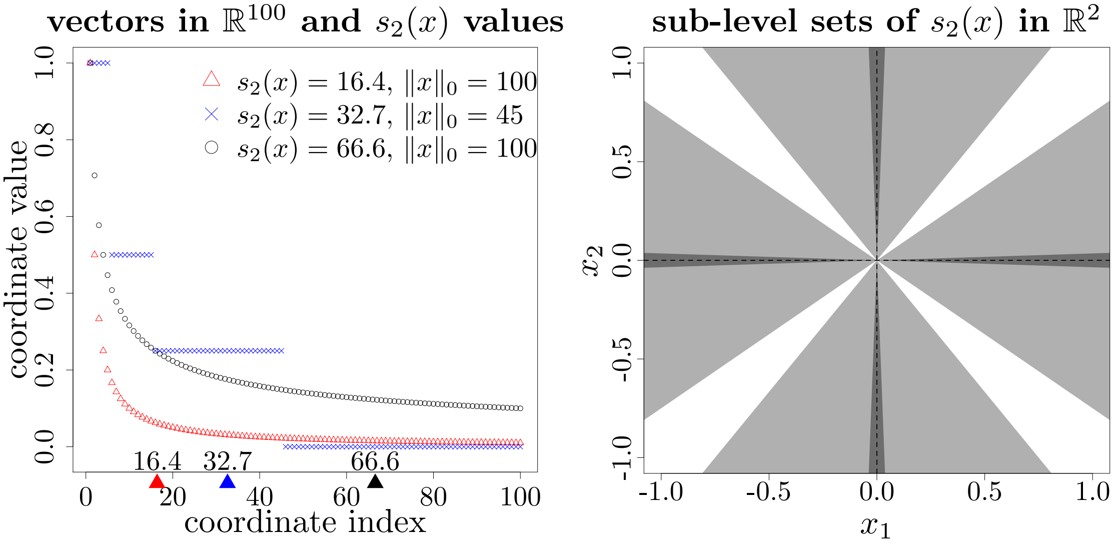

Figure 3.1: Characteristics of $s_2(x)$. Left panel: Three vectors (red, blue, black) in $\mathbb{R}^{100}$ have been plotted with their coordinates in order of decreasing size (maximum entry normalized to 1). Two of the vectors have power-law decay profiles, and one is a dyadic vector with exactly 45 positive coordinates (red: $x_i \propto i^{-1}$, blue: dyadic, black: $x_i \propto i^{-1/2}$). Color-coded triangles on the bottom axis indicate that the $s_2(x)$ value represents the "effective" number of coordinates. Right panel: The light grey set is given by $\{x \in \mathbb{R}^2 : s_2(x) \leq 1.9\}$, and the dark grey set is given by $\{x \in \mathbb{R}^2 : s_2(x) \leq 1.1\}$.

Numerically stable measures of rank and sparsity for matrices

The framework of CS naturally extends to the problem of recovering an unknown matrix $X \in \mathbb{R}^{p_1 \times p_2}$ on the basis of the measurement model

$$y = \mathcal{A}(X) + \sigma\epsilon, \tag{3.14}$$

where $y \in \mathbb{R}^n$, and $\mathcal{A}$ is a user-specified linear operator from $\mathbb{R}^{p_1 \times p_2}$ to $\mathbb{R}^n$. In recent years, many researchers have explored the recovery of X when it is assumed to have sparse or low rank structure. We refer to the papers [CP11; Cha+12] for descriptions of numerous applications. In analogy with the previous section, the parameters $\text{rank}(X)$ or $\|X\|_0$ play important theoretical roles, but are very sensitive to perturbations of X. Likewise, it is of basic interest to estimate robust measures of rank and sparsity for matrices. Since the analogue of $s_q(\cdot)$ for measuring matrix sparsity is easily derived by viewing X as a vector in $\mathbb{R}^{p_1 p_2}$, we restrict our attention to the more distinct issue of soft measures of rank.

In the context of recovering a low-rank matrix X the quantity $\text{rank}(X)$ plays the role that the norm $\|x\|_0$ does in the recovery of a sparse vector. If we let $\varsigma(X) \in \mathbb{R}^p_+$ denote the vector of ordered singular values of X, the connection can be made explicit by writing

$$\text{rank}(X) = \|\varsigma(X)\|_0.$$

As in our discussion of sparsity, it is of basic interest to consider a numerically stable version of the usual rank function. Motivated by the definition of $s_q(x)$ in the vector case, we can also consider

$$r_q(X) := s_q(\varsigma(X)) = \left(\frac{\|\varsigma(X)\|_q}{\|\varsigma(X)\|_1}\right)^{\frac{q}{1-q}} = \left(\frac{|||X|||_q}{|||X|||_1}\right)^{\frac{q}{1-q}}$$

as a measure of the effective rank of X, where $q > 0$ and $|||X|||_q := \|\varsigma(X)\|_q$. (When $q \geq 1$, $|||X|||_q$ is known as the Schatten q-norm of X.) Essentially all of the properties of $s_q(\cdot)$ described earlier carry over to $r_q(\cdot)$ in a natural way, and so we do not state these in detail. We also note that quantities related to $r_q(X)$, or special instances, have been considered elsewhere as a measure of rank, e.g. the *numerical rank*[5] $\|X\|_F^2/\|X\|_{\text{op}}^2$ [RV07], or the instance $r_2(X)$ [LJW11; TN12; NW12].

Contributions

The main contributions of the chapter can be summarized in three parts.

The family of sparsity measures $\{s_q(x)\}_{q\geq 0}$. As mentioned in Section 3.1, our definition of $s_q(x)$ in terms of Rényi entropy gives a conceptual foundation for several norm ratios that have appeared elsewhere in the sparsity literature. Furthermore, we clarify the meaning of $s_2(x)$ with regard to signal recovery by showing in Section 3.2 that $s_2(x)$ plays an intuitive role in the performance of the Basis Pursuit Denoising (BPDN) algorithm. Specifically, we show that the relative ℓ_2 error of BPDN can be bounded in a sharp way by the quantity $\sqrt{s_2(x)\log(ep/n)/n}$, which is formally similar to the well-known rate of ℓ_2 approximation $\sqrt{k\log(p)/n}$ for k-sparse signals [Neg+12]. This connection is explained more carefully in Section 3.2.

Estimation results, confidence intervals, and applications. Our central methodological contribution is a new deconvolution-based approach for estimating $\|x\|_q$ and $s_q(x)$ from linear measurements. The procedure we propose is of particular interest in the way that it blends the tools of *sketching with stable laws* and *deconvolution with characteristic*

[5] This can be cast in the framework of $r_q(X)$ by defining the probability vector $\pi(x)$ as $\pi_j(x) = |x_j|^2/\|x\|_2^2$ in the definition of $s_q(x)$ and then choosing $q = \infty$.

functions. These tools are typically applied in different contexts, as discussed in Section 3.1. Also, the computational cost of our procedure is small relative to the cost of recovering the full signal by standard methods.

In terms of consistency, the most important features of our estimator $\widehat{s}_q(x)$ are that it does not rely on any sparsity assumptions, and that its relative error converges to 0 at the *dimension-free* rate of $1/\sqrt{n}$ (in probability). Consequently, only $\mathcal{O}(1)$ measurements are needed to obtain a good estimate of $s_q(x)$, even when p is large. As explained in Section 3.1, this result naturally suggests a two-stage measurement process: First, a small initial measurement price can be paid to obtain $\widehat{s}_q(x)$. Second, the value $\widehat{s}_q(x)$ can be used to adaptively select "just enough" extra measurements for recovering the full signal. (Proposition 3.1 in Section 3.2 indicates that this number can be chosen proportionally to $\widehat{s}_2(x)\log(p)$ when BPDN is used for recovery.)

In addition to proving ratio-consistency, we derive a CLT for $\widehat{s}_q(x)$, which allow us to obtain confidence intervals $s_q(x)$ with asymptotically exact coverage probability. A notable feature of this CLT is that it is "uniform" with respect to the tuning parameter in our procedure for computing $\widehat{s}_q(x)$. The uniformity is important because it allows us to make an optimal *data-dependent* selection of the tuning parameter and still find the estimator's limiting distribution (see Theorem 3.2 and Corollary 3.1). In terms of applications, we show in Section 3.5 how this CLT can be used in inferential problems related to unknown sparsity, i.e. testing the null hypothesis that $s_q(x)$ is greater than a given level, and ensuring that the true signal lies in the constraint set of the (primal) Lasso or Elastic net problems with a given degree of statistical significance.

The necessity of randomized measurements. At the present time, the problem of constructing deterministic measurement matrices with performance guarantees comparable to those of random matrices is one of the major unresolved theoretical issues in CS [FR13, Section 1.3] [CHJ10]. Due to the fact that our proposed method for estimating $s_q(x)$ depends on randomized measurements, one may similarly wonder if randomization is essential to the problem of estimating unknown sparsity. In Section 3.6, we show that randomization is essential from a worst-case point of view. Our main result in this direction (Theorem 3.4) shows that for any deterministic matrix $A \in \mathbb{R}^{n\times p}$, and any deterministic procedure for estimating $s_q(x)$, there is always at least one signal for which the relative estimation error is at least of order 1, even if the measurements are noiseless. This contrasts with performance our randomized method, whose relative error is $\mathcal{O}_P(1/\sqrt{n})$ for any choice of x. Furthermore, the result has a negative implication for sparse linear regression. Namely, due to the fact that the design matrix is often viewed as "fixed and given" in many regression problems, our result indicates that $s_q(x)$ cannot be consistently estimated in relative error in that context (from a worst-case point of view).

Related work

Our work here substantially extends our earlier conference paper [Lop13] and has connections with a few different lines of research.

Extensions beyond the conference paper [Lop13]. Whereas our earlier work deals exclusively with the sparsity measure $\|x\|_1^2/\|x\|_2^2$, the current chapter considers the estimation of the family of parameters $s_q(x)$. The procedure we propose for estimating $s_q(x)$ (as well as our analysis of its performance) include several improvements on the earlier approach. In particular, the new procedure tolerates noise with infinite variance and leads to confidence intervals with asymptotically exact coverage probability (whereas the previous approach led to conservative intervals). Also, the applications of our procedure to tuning recovery algorithms and testing the hypothesis of sparsity are new (Section 3.5). Lastly, our theoretical results in Sections 3.2 and 3.6 may be regarded as sharpened versions of parallel results in the paper [Lop13].

Sketching with stable laws. Our approach to estimating $s_q(x)$ is based on the subproblem of estimating $\|x\|_q$ for various choices of q. In order to estimate such norms from linear measurements, we employ the technique of *sketching with stable laws*, which has been developed extensively in the streaming computation literature. (The book [Cor+12] offers an overview, and seminal papers include [Ind06] and [AMS96]). Over the last few years, the exchange of ideas between sketching and CS has just begun to accelerate, as in the papers [GI10] [Lop13] [LZZ14] [Ind13]. Nevertheless, to the best of our knowledge, the present chapter and our earlier work [Lop13] are the first to apply sketching ideas to the problem of unknown sparsity in CS.

In essence, our use of the sketching technique is based on the fact that if a random vector $a_1 \in \mathbb{R}^p$ has i.i.d. coordinates drawn from a standard symmetric q-stable law,[6] then the random variable $\langle a_1, x\rangle$ has a q-stable law whose scale parameter is equal to $\|x\|_q$. (A more detailed introduction is given in Section 3.3.) Consequently, the problem of estimating $\|x\|_q$ can be thought of as estimating the scale parameter of a stable law convolved with noise.

In the streaming computation literature, the observation model giving rise to $\langle a_1, x\rangle$ is quite different than in CS. Roughly speaking, the vector x is thought of as a massive data stream whose entries can be observed sequentially, but cannot be stored entirely in memory. The core idea is that by computing "sketches" $\langle a_1, x\rangle = a_{11}x_11 + a_{12}x_2 + \cdots$ in a sequential manner, it is possible to estimate various functions of x from the sketches without having to store the entire stream. Under this framework, a substantial body of work has studied the estimation of ℓ_q norms with $q \geq 0$ [CC12] [Cor+03] [Li08] [LHC07] [Ind06] [AMS96] [Fei+02]. However, results in this direction are typically not directly applicable to CS, due to essential differences in the observation model. For instance, measurement noise does not generally play a role in the sketching literature.

[6] e.g. Gaussian when $q = 2$ and Cauchy when $q = 1$.

Empirical characteristic functions. As just mentioned, our approach to estimating $\|x\|_q$ and $s_q(x)$ can be thought of as deconvolving the scale parameter of a stable law. Given that stable laws have a simple analytic formula for their characteristic function (and have no general formula for their likelihood function), it is natural to use the empirical characteristic function $\widehat{\Psi}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \exp(\sqrt{-1}ty_i)$ as a foundation for our estimation procedure. With regard to denoising, characteristic function are also attractive insofar as they factors over convolution, and exists even when the noise distribution is heavy-tailed.

The favorable properties of empirical characteristic functions have been applied by several authors to the deconvolution of scale parameters [MHL95] [MH95], [B+05] [Mei06] [Mat02]. Although the basic approach used in these papers is similar to ours, the results in these works are not directly comparable with ours due to differences in model assumptions. Another significantly distinct aspect of our work deals with the choice of the "tuning parameter" t in the function $\widehat{\Psi}_n(t)$. This choice is a basic element in most methods based on empirical characteristic functions. In detail, we show how to make an optimal data-adaptive choice $\widehat{t}$, and we derive the limiting distribution of the estimator $\widehat{s}_q(x)$ that originates from $\widehat{\Psi}_n(\widehat{t})$. This leads to a significant technical challenge in accounting for the randomness of $\widehat{t}$, and in order to do this, we show that the process $\widehat{\Psi}_n(\cdot)$ arising from our model assumptions satisfies a uniform CLT in the space $\mathscr{C}(\mathcal{I})$ of continuous complex functions on a compact interval $\mathcal{I}$. (See the paper [Mar81] or the book [Ush99] for more details concerning weak convergence of empirical characteristic functions.) With regard to the cited line of works concerning deconvolution of scale parameters, it seems that our work is the first to derive the limiting distribution of the scale estimator under a data-dependent choice of tuning parameter.

Model selection and validation in CS. Some of the challenges described in Section 3.1 can be approached with the general tools of cross-validation (CV) and empirical risk minimization (ERM). This approach has been used to select various parameters in CS, such as the number of measurements n [MSW08; War09], the number of OMP iterations k [War09], or the Lasso regularization parameter λ [Eld09]. At a high level, these methods consider a collection of (say m) solutions $\widehat{x}^{(1)}, \dots, \widehat{x}^{(m)}$ obtained from different values $\theta_1, \dots, \theta_m$ of some tuning parameter of interest. For each solution, an empirical error estimate $\widehat{\text{err}}(\widehat{x}^{(j)})$ is computed, and the value θ_{j^*} corresponding to the smallest $\widehat{\text{err}}(\widehat{x}^{(j)})$ is chosen.

Although methods based on CV/ERM share common motivations with our work here, these methods differ from our approach in several ways. In particular, the problem of estimating a soft measure of sparsity, such as $s_q(x)$, has not been considered from that angle. Also, the cited methods do not give any theoretical guarantees to ensure that the chosen tuning parameter leads to a solution whose ℓ_0 sparsity level is close to the true one. (Note that even if CV suggests that an estimate $\widehat{x}$ has small error $\|\widehat{x}-x\|_2$, it is not necessary for $\|\widehat{x}\|_0$ to be close to $\|x\|_0$.) This point is especially relevant in inferential problems, such as identifying a set of important variables or making confidence statements related to an unknown sparsity value. From a computational point view, the CV/ERM approaches can also be costly — since $\widehat{x}^{(j)}$ may need to be computed from a separate optimization problem for for each choice of the tuning parameter. By contrast, our method for estimating $s_q(x)$ requires very little computation.

Outline

The remainder of the chapter is organized as follows. In Section 3.2, we formulate a (tight) recovery guarantee for the Basis Pursuit Denoising algorithm directly in terms of $s_2(x)$. Next, in Section 3.3, we propose estimators for $\|x\|_q$ and $\widehat{s}_q(x)$, and in Section 3.4 we state consistency results and provide confidence intervals for $\|x\|_q$ and $s_q(x)$. Applications to testing the hypothesis of sparsity and adaptive tuning of the Lasso are presented in Section 3.5. In Section 3.6, we show that the use of randomized measurements is essential to estimating $s(x)$ in a minimax sense. We defer all of the proofs to Appendix B.

3.2 Recovery guarantees in terms of $s_2(x)$

In this section, we state two simple propositions that illustrate the link between $s_2(x)$ and recovery conditions for the Basis Pursuit Denoising (BPDN) algorithm [CDS98]. The main purpose of these results is to highlight the fact that $s_2(x)$ and $\|x\|_0$ play analogous roles with respect to the sample complexity of sparse recovery. Specifically, we provide *matching* upper and lower bounds for relative ℓ_2 reconstruction error of BPDN in terms of $s_2(x)$. These bounds also suggest two applications of the quantity $s_2(x)$. First, the order of the reconstruction error can be estimated whenever $s_2(x)$ can be estimated. Second, when measurements can be collected sequentially, an estimate of $s_2(x)$ from an initial set of measurements allows for the user to select a number of secondary measurements that adapts to the particular structure of x, e.g. $n = \widehat{s}_2(x)\log(p)$.

Setup for BPDN. In order to explain the connection between $s_2(x)$ and recovery, we first recall a fundamental result describing the ℓ_2 error rate of the BPDN algorithm. Here, it will be convenient to combine Theorems **??** and 1.4 stated earlier in Section **??** of Chapter 1. For the first two results of this section, we will work under two standard assumptions underlying those theorems.

A3.1. *There is a constant ϵ_0 such that all realizations of the noise vector $\epsilon \in \mathbb{R}^n$ satisfy $\|\epsilon\|_2 \leq \epsilon_0$.*

A3.2. *The entries of $A \in \mathbb{R}^{n\times p}$ are an i.i.d. sample from $\frac{1}{\sqrt{n}}G_0$, where G_0 is a sub-Gaussian distribution with mean 0 and variance 1.*

Since ϵ and A are both random, probability statements will be made with respect to their joint distribution. When the noise distribution satisfies **A3.1**, the output of the BPDN algorithm is a solution to the following convex optimization problem

$$\widehat{x} \in \operatorname{argmin}\big\{\|v\|_1 : \|Av - y\|_2 \leq \sigma\epsilon_0, v \in \mathbb{R}^p\big\}. \tag{BPDN}$$

As a final piece of notation, for any $T \in \{1, \dots, p\}$, we use $x_{|T}$ to denote the best T-term approximation of x, which is computed by retaining the largest T entries of x in magnitude, and setting all others to 0.

Theorem 3.1 ([CRT06; CWX10; Ver12])**.** *Suppose the model* (3.1) *satisfies the conditions **A3.1** and **A3.2**. Let $x \in \mathbb{R}^p$ be arbitrary, and fix a number $T \in \{1, \dots, p\}$. Then, there are absolute constants $c_2, c_3 > 0$, and numbers $c_0, c_1 > 0$ depending only on the distribution G_0, such that the following statement is true. If*

$$n \geq c_0 T \log(pe/T), \tag{3.15}$$

then with probability at least $1-2\exp(-c_1 n)$, any solution $\widehat{x}$ to the problem (BPDN) *satisfies*

$$\|\widehat{x} - x\|_2 \leq c_2\, \sigma\epsilon_0 + c_3\, \tfrac{\|x - x_{|T}\|_1}{\sqrt{T}}. \tag{3.16}$$

An upper bound in terms of $s_2(x)$. Two important aspects of Theorem 3.1 are that it holds for *all* signals $x \in \mathbb{R}^p$, and that it measures sparsity via the T-term approximation error $\|x - x_{|T}\|_1$, rather than the idealized ℓ_0 norm. However, a main limitation is that the detailed relationship between T and the approximation error $\frac{1}{\sqrt{T}}\|x - x_{|T}\|_1$ is typically unknown for the true signal x. Consequently, it is not clear how large n should be chosen in line (3.15) to ensure that $\|x - \widehat{x}\|_2$ is small with high probability. The next proposition resolves this issue by modifying the bound (3.16) so that that the relative ℓ_2 error is bounded by an explicit function of n and the estimable parameter $s_2(x)$.

Proposition 3.1. *Assume conditions **A3.1** and **A3.2** hold, and let $x \in \mathbb{R}^p \backslash \{0\}$ be arbitrary. Then, there is an absolute constant $c_2 > 0$, and numbers $c_1, c_3 > 0$ depending only on the distribution G_0, such that the following statement is true. If n and p satisfy $\log(\frac{pe}{n}) \leq n \leq p$, then with probability at least $1-2\exp(-c_1 n)$, any solution $\widehat{x}$ to the problem* (BPDN) *satisfies*

$$\tfrac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq c_2 \tfrac{\sigma\epsilon_0}{\|x\|_2} + c_3 \sqrt{\tfrac{s_2(x)\log(\frac{pe}{n})}{n}}. \tag{3.17}$$

Remarks. Note that this result holds for any n of modest size, $n \geq \log(pe/n)$. The bound also has a familiar form in relation to other well known recovery guarantees for hard sparse signals, with $s_2(x)$ playing a role that is similar to $\|x\|_0$.

To illustrate the connection between $s_2(x)$ and $\|x\|_0$ in greater detail, we now briefly summarize a standard bound on the relative ℓ_2 error of the Lasso in recovering hard sparse signals. The Lasso estimator is defined by

$$\widehat{x}_\lambda \in \text{argmin}\{\tfrac{1}{2}\|y - Av\|_2^2 + \lambda\|v\|_1 : v \in \mathbb{R}^p\}. \tag{LASSO}$$

Suppose we observe $y = Ax + \sigma\epsilon$, where $\epsilon \sim N(0, \frac{1}{n}I_{n\times n})$, and the number of measurements is at least of order $n \gtrsim \|x\|_0 \log(p)$. In addition, suppose that the matrix A is constructed as $\frac{1}{\sqrt{n}}Z \in \mathbb{R}^{n\times p}$ where the entries of Z are i.i.d. Rademacher variables (±1 with equal probability). Then, when the regularization parameter is set to $\lambda = 4\sigma\sqrt{\log(p)/n}$, a Lasso solution $\widehat{x}_\lambda$ will satisfy the bound

$$\tfrac{\|\widehat{x}_\lambda - x\|_2}{\|x\|_2} \le c\big(\tfrac{\sigma}{\|x\|_2}\big)\sqrt{\tfrac{\|x\|_0 \log(p)}{n}} \tag{3.18}$$

with high probability, where c is an absolute constant. (See the paper [Neg+12] for additional details and a more general statement of the result.[7] Although the bounds (3.17) and (3.18) rely on strictly different assumptions regarding the noise vector, the assumptions are qualitatively similar in certain situations. Note that when $\epsilon \sim N(0, \frac{1}{n}I_{n\times n})$, the norm $\|\epsilon\|_2$ is close to 1 with high probability, and so the assumption that $\|\epsilon\|_2 \le \epsilon_0$ underlying line (3.17) is comparable if ϵ_0 is close to 1. To compare the bounds (3.17) and (3.18), observe that if n is taken to grow at the "minimally sufficient" rate $n \asymp \|x\|_0 \log(p)$, then the fact that $s_2(x) \le \|x\|_0$ implies

$$\sqrt{\tfrac{s_2(x)\log(pe/n)}{n}} \lesssim 1. \tag{3.19}$$

Under this scaling of n, if we further assume that the noise to signal ratio $\sigma/\|x\|_2$ is non-vanishing (e.g. when the signal has bounded energy and σ is held fixed), then it follows that both of the bounds (3.17) and (3.18) are of order 1.

A matching lower bound in terms of $s_2(x)$. Our next result shows that the upper bound (3.17) is sharp in the case of noiseless measurements. More precisely, for any choice of $A \in \mathbb{R}^{n\times p}$, there is always at least one signal $\tilde{x} \in \mathbb{R}^p \setminus \{0\}$ for which the relative ℓ_2 error of BPDN is at least $\sqrt{s_2(\tilde{x})\log(pe/n)/n}$, up to an absolute constant. In fact, the lower bound is applicable beyond BPDN, and imposes a limit of performance on all algorithms that satisfy the mild condition of being *homogenous* in the noiseless setting. To be specific, if a recovery algorithm is viewed as a map $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^p$ that sends a vector of noiseless measurements $Ax \in \mathbb{R}^n$ to a solution $\widehat{x} = \mathcal{R}(Ax) \in \mathbb{R}^p$, then $\mathcal{R}$ is said to be homogenous if

$$\mathcal{R}(A(cx)) = c \cdot \mathcal{R}(Ax) \quad \text{for all } c > 0. \tag{3.20}$$

It is simple to verify that the BPDN is homogenous in the case of noiseless measurements, since it reduces to the ordinary Basis Pursuit (BP) algorithm, whose solution may be written as[8]

$$\widehat{x} \in \operatorname{argmin}\{\|v\|_1 : Av = y, v \in \mathbb{R}^p\}. \tag{BP}$$

[7] Although the paper [Neg+12] does not present their result (Corollary 2) with the noise variables having variance of order $1/n$, we use this scaling when quoting their result so that it can be compared on equal footing with the BPDN result in Theorem 3.1. The difference in scaling arises from the fact that the entries of A are roughly of order $1/\sqrt{n}$ in the paper [CRT06], whereas the entries of A are roughly of order 1 in the paper [Neg+12]. Note also that in our definition of $\widehat{x}_\lambda$ the first term of the objective function is $\frac{1}{2}\|Av - y\|_2^2$ whereas in the paper [Neg+12] it is $\frac{1}{2n}\|Av - y\|_2^2$.

[8] If this minimization problem does not have a unique optimal solution, we still may still regard BPDN as a well defined function from $\mathbb{R}^n$ to $\mathbb{R}^p$ by considering a numerical implementation that never returns more than one output for a given input.

Apart from the basic condition of homogeneity, our lower bound requires no other assumptions. Note also that the statement of the result does not involve any randomness.

Proposition 3.2. *There is an absolute constant $c_0 > 0$ for which the following statement is true. For any homogenous recovery algorithm $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^p$, and any $A \in \mathbb{R}^{n\times p}$ with $n \leq p$, there is at least one point $\tilde{x} \in \mathbb{R}^p \setminus \{0\}$ such that*

$$\tfrac{\|\widehat{x}-\tilde{x}\|_2}{\|\tilde{x}\|_2} \geq c_0\sqrt{\tfrac{s_2(\tilde{x})\log(\frac{pe}{n})}{n}}, \tag{3.21}$$

where $\widehat{x} = \mathcal{R}(A\tilde{x})$.

3.3 Estimation procedures for $s_q(x)$ and $\|x\|_q^q$

In this section, we describe a procedure to estimate $s_q(x)$ for an arbitrary non-zero signal x, and any $q \in (0, 2] \setminus \{1\}$. The procedure uses a small number of measurements, makes no sparsity assumptions, and requires very little computation. The measurements we prescribe may also be re-used to recover the full signal after the parameter $s_q(x)$ has been estimated. In the the first three subsections, we present the ideas underlying the procedure, and then in the last subsection, we describe the procedure as an algorithm.

The deconvolution model

Here we describe the model assumptions that our estimation procedure for $s_q(x)$ will be based on. (These are different from the assumptions used in the previous section.) In scalar notation, we consider linear measurements given by

$$y_i = \langle a_i, x\rangle + \sigma\epsilon_i, \qquad i = 1, \ldots, n. \tag{M}$$

Model assumptions. For the remainder of the paper, we assume $x \neq 0$ unless stated otherwise. Regarding the noise variables ϵ_i, we assume they are generated in an i.i.d. manner from a distribution denoted by F_0. When the a_i are generated randomly, we assume that $\{a_1, \ldots, a_n\}$ is an independent set of random vectors, and also that the sets $\{a_1, \ldots, a_n\}$ and $\{\epsilon_1, \ldots, \epsilon_n\}$ are independent. The noise variables are assumed to be symmetric about 0 and to satisfy $0 < \mathbb{E}|\epsilon_1| < \infty$, but they may have *infinite variance.* A minor technical condition we place on F_0 is that the roots of its characteristic function φ_0 are isolated. This condition is satisfied by a broad range of naturally occurring distributions, and in fact, many works on deconvolution assume that φ_0 has no roots at all.[9] The noise scale parameter $\sigma > 0$ and the distribution F_0 are treated as being known, which is a common assumption in deconvolution problems. Also note that in certain situations, it may be possible to directly estimate F_0 by using "blank measurements" with $a_i = 0$.

[9] It is known that a subset of $\mathbb{R}$ is the zero set of a characteristic function if and only if it symmetric, closed, and excludes 0. [Ili76; Gne01].

Asymptotics. Following the usual convention of high-dimensional asymptotics, we allow the model parameters to vary as $(n,p) \to \infty$. This means that there is an implicit index $\xi \in \mathbb{Z}_+$, such that $n = n(\xi)$, $p = p(\xi)$ and both diverge as $\xi \to \infty$. It will turn out that our asymptotic results will not depend on the ratio p/n, and so we allow p to be arbitrarily large with respect to n. We also allow $x = x(\xi)$, $\sigma = \sigma(\xi)$ and $a_i = a_i(\xi)$, but the noise distribution F_0 is fixed with respect to ξ. When making asymptotic statements about probability, we view the set of pairs $\{(a_1, \epsilon_1), \dots, (a_n, \epsilon_n)\}$ as forming a triangular array with rows indexed by ξ, and columns indexed by $n(\xi)$. Going forward, we will generally suppress the index ξ.

Sketching with stable laws in the presence of noise

For any $q \in (0,2]$, the sketching technique offers a way to estimate $\|x\|_q^q$ from a set of randomized linear measurements. Building on this technique, we estimate $s_q(x) = (\|x\|_q/\|x\|_1)^{q/(1-q)}$ by estimating $\|x\|_q^q$ and $\|x\|_1$ from separate sets of measurements. The core idea is to generate the measurement vectors $a_i \in \mathbb{R}^p$ using *stable laws*. A standard reference on this class of distributions is the book [Zol86].

Definition 1. A random variable V has a *symmetric q-stable distribution* if its characteristic function is of the form $\mathbb{E}[\exp(\sqrt{-1}tV)] = \exp(-|\gamma t|^q)$ for some $q \in (0,2]$ and some $\gamma > 0$, where $t \in \mathbb{R}$. We denote the distribution by $V \sim \text{stable}_q(\gamma)$, and γ is referred to as the *scale* parameter.

The most well-known examples of symmetric stable laws are the cases of $q = 2$ and $q = 1$. Namely, $\text{stable}_2(\gamma)$ is the Gaussian distribution $N(0, 2\gamma^2)$, and $\text{stable}_1(\gamma)$ is the Cauchy distribution $C(0,\gamma)$. To fix some notation, if a vector $a_1 = (a_{11}, \dots, a_{1p}) \in \mathbb{R}^p$ has i.i.d. entries drawn from $\text{stable}_q(\gamma)$, we write $a_1 \sim \text{stable}_q(\gamma)^{\otimes p}$. Also, since our work will involve different choices of q, we will write γ_q instead of γ from now on. The connection with ℓ_q norms hinges on the following property of stable distributions, which is simple to derive from Definition 1.

Lemma 3.1. *Suppose $x \in \mathbb{R}^p$ is fixed, and $a_1 \sim \text{stable}_q(\gamma_q)^{\otimes p}$ with parameters $q \in (0,2]$ and $\gamma_q > 0$. Then, the random variable $\langle x, a_1 \rangle$ is distributed according to* $\text{stable}_q(\gamma_q \|x\|_q)$.

Using this fact, if we generate a set of i.i.d. measurement vectors $a_1, \dots, a_n$ from the distribution $\text{stable}_q(\gamma_q)^{\otimes p}$ and let $\tilde{y}_i = \langle a_i, x \rangle$, then $\tilde{y}_1, \dots, \tilde{y}_n$ is an i.i.d. sample from $\text{stable}_q(\gamma_q \|x\|_q)$. Hence, in the special case of noiseless linear measurements, the task of estimating $\|x\|_q^q$ is equivalent to a well-studied univariate problem: *estimating the scale parameter of a stable law from an i.i.d. sample.* In the presence of noise, this key idea can be substantially extended to obtain a consistent estimator of $\|x\|_q^q$, as is shown in Section 3.4. We emphasize that the extra deconvolution step is an important aspect of our method that distinguishes it from existing work in the sketching literature (where noise is typically not considered).

When generating the measurement vectors from $\text{stable}_q(\gamma_q)$, the parameter γ_q governs the "energy level" of the a_i. For instance, in the case of Gaussian measurements with $a_i \sim \text{stable}_2(\gamma_2)^{\otimes p}$, we have $\mathbb{E}\|a_i\|_2^2 = 2p\gamma_2^2$. In general, for any q, as the energy level is increased, the effect of noise is diminished. Likewise, in our analysis, we view γ_q as a physical aspect of the measurement system that is known to the user. Asymptotically, we allow $\gamma_q = \gamma_q(\xi)$ to vary as $(n,p) \to \infty$ in order to reveal the trade-off between the measurement energy and the noise level σ.

The sub-problem of estimating $\|x\|_q^q$

Our procedure for estimating $s_q(x)$ uses two separate sets of measurements of the form (M) to compute estimators $\widehat{\|x\|_1}$ and $\widehat{\|x\|_q^q}$. The respective sizes of each measurement set will be denoted by n_1 and n_q. To unify the discussion, we will describe just one procedure to compute $\widehat{\|x\|_q^q}$ for any $q \in (0,2]$, since the ℓ_1-norm estimator is a special case. The two estimators are then combined to obtain an estimator of $s_q(x)$, defined by

$$\widehat{s}_q(x) := \frac{\left(\widehat{\|x\|_q^q}\right)^{\frac{1}{1-q}}}{\left(\widehat{\|x\|_1}\right)^{\frac{q}{1-q}}}, \tag{3.22}$$

which makes sense for any $q \in (0,2]$ except $q = 1$. Of course, the parameters $\|x\|_0$ and $s_1(x)$ can still be estimated in practice by using $\widehat{s}_q(x)$ for some value of q that is close to 0 or 1. Indeed, the ability to approximate $\|x\|_0$ and $s_1(x)$ in this way is a basic motivation for studying $s_q(x)$ over a continuous range of q.

An estimating equation based on characteristic functions

Characteristic functions offer a very natural route toward estimating $\|x\|_q^q$. If we draw i.i.d. measurement vectors

$$a_i \sim \text{stable}_q(\gamma_q)^{\otimes p}, \qquad i = 1, \ldots, n_q,$$

then the characteristic function of the measurement $y_i = \langle a_i, x\rangle + \sigma\epsilon_i$ is given by

$$\Psi(t) := \mathbb{E}[\exp(\sqrt{-1}ty_i)] = \exp(-\gamma_q^q |t|^q \|x\|_q^q) \cdot \varphi_0(\sigma t), \tag{3.23}$$

where $t \in \mathbb{R}$, and we recall that φ_0 denotes the characteristic function of the noise variables ϵ_i. Note that φ_0 is real-valued since we assume that the noise-distribution is symmetric about 0. Using the measurements $y_1, \ldots, y_{n_q}$, we can approximate $\Psi(t)$ by computing the empirical characteristic function

$$\widehat{\Psi}_{n_q}(t) := \frac{1}{n_q}\sum_{i=1}^{n_q} e^{\sqrt{-1}ty_i}. \tag{3.24}$$

Next, by solving for $\|x\|_q^q$ in the approximate equation

$$\widehat{\Psi}_{n_q}(t) \approx \exp(-\gamma_q^q |t|^q \|x\|_q^q) \cdot \varphi_0(\sigma t), \tag{3.25}$$

we obtain an estimator $\widehat{\|x\|_q^q}$. To make the dependence on t explicit, we will mostly use the notation $\widehat{\nu}_q(t) = \widehat{\|x\|_q^q}$. Proceeding with the arithmetic in the previous line leads us to define

$$\widehat{\nu}_q(t) := \tfrac{-1}{\gamma_q^q |t|^q} \operatorname{Log}_+ \operatorname{Re}\Big(\tfrac{\widehat{\Psi}_{n_q}(t)}{\varphi_0(\sigma t)}\Big), \tag{3.26}$$

when $t \neq 0$ and $\varphi_0(\sigma t) \neq 0$. Here, the symbol $\operatorname{Re}(z)$ denotes the real part of a complex number z. Also, we define $\operatorname{Log}_+(r) := \log(|r|)$ for any real number $r \neq 0$, and $\operatorname{Log}_+(0) := 1$. For the particular values of t where $t = 0$ or $\varphi_0(\sigma t) = 0$, we arbitrarily define $\widehat{\nu}_q(t) = 1$. The need to use Log_+ and handle these particular values of t will be irrelevant from an asymptotic point of view. We only mention these details for the technical convenience of having an estimator that is defined for all values of $t \in \mathbb{R}$.

Optimal selection of the tuning parameter

A crucial aspect of the estimator $\widehat{\nu}_q(t)$ is the choice of $t \in \mathbb{R}$, which plays the role of a tuning parameter. This choice turns out to be somewhat delicate, especially in situations where $\|x\|_q \to \infty$ as $(n, p) \to \infty$. To see why this matters, consider the equation

$$\tfrac{\widehat{\nu}_q(t)}{\|x\|_q^q} = \tfrac{-1}{(\gamma_q |t| \|x\|_q)^q} \operatorname{Log}_+ \operatorname{Re}\Big(\tfrac{\widehat{\Psi}_{n_q}(t)}{\varphi_0(\sigma t)}\Big). \tag{3.27}$$

If we are in a situation where $\|x\|_q$ diverges while the parameters γ_q, σ, and t remain of order 1, then the empirical charcteristic function $\widehat{\Psi}_{n_q}(t)$ will collapse to 0 (due to line (3.25). The right hand side of line (3.27) may then become unstable as it can tend to a limit of the form $\frac{\infty}{\infty}$. Hence, it is desirable to choose t adaptively so that as $(n, p) \to \infty$,

$$\gamma_q t \|x\|_q \to c_0, \tag{3.28}$$

for some finite constant $c_0 > 0$, which prevents $\widehat{\Psi}_{n_q}(t)$ from collapsing to 0. When this desired scaling can be achieved, the next step is to further refine choice of t so as to minimize the limiting variance of $\widehat{\nu}_q(t)$. Our proposed method will solve both of these problems.

Of course, the ability to choose t adaptively requires some knowledge of $\|x\|_q$, which is precisely the quantity we are trying to estimate! As soon as we select a data-dependent value, say $\widehat{t}$, we introduce a significant technical challenge: Inferences based on the adaptive estimator $\widehat{\nu}_q(\widehat{t}\,)$ must take the randomness in $\widehat{t}$ into account. Our approach is to prove a uniform CLT for the function $\widehat{\nu}_q(\cdot)$. As will be shown in the next result, the uniformity will allow us to determine the limiting law of $\widehat{\nu}_q(\widehat{t}\,)$ as if the optimal choice of t was known in advance of observing any data. To make the notion of optimality precise, we first describe the limiting law of $\widehat{\nu}_q(\widehat{t})$ for any data-dependent value $\widehat{t}$ that satisfies the scaling condition (3.28) (in probability). A method for constructing such a value $\widehat{t}$ will be given the next subsection. Consistency results for the procedure are given in Section 3.4.

To state the uniform CLT, we need to introduce the *noise-to-signal* ratio

$$\rho_q = \rho_q(\xi) := \tfrac{\sigma}{\gamma_q \|x\|_q}, \tag{3.29}$$

which will be related to the width of our confidence interval for $\|x\|_q^q$. Although we allow ρ_q to vary with (n_q, p), we will assume that it stabilizes to a finite limiting value:

A3.3. *For each $q \in (0, 2]$, there is a limiting constant $\bar{\rho}_q \in [0, 1]$ such that $\rho_q = \bar{\rho}_q + o(n^{-1/2})$ as $(n_q, p) \to \infty$.*

This assumption merely encodes the idea that the signal is not overwhelmed by noise asymptotically.

Theorem 3.2 (Uniform CLT for ℓ_q norm estimator)**.** *Let $q \in (0, 2]$. Assume that the measurement model* (M) *and Assumption 3.3 hold. Let $\widehat{t}$ be any function of $y_1, \dots, y_{n_q}$ that satisfies*

$$\widehat{t}\gamma_q \|x\|_q \longrightarrow_P c_0 \tag{3.30}$$

as $(n_q, p) \to \infty$ for some constant $c_0 \neq 0$ with $\varphi_0(\bar{\rho}_q c_0) \neq 0$. Then, the estimator $\widehat{\nu}_q(\widehat{t})$ satisfies

$$\sqrt{n_q}\Big(\tfrac{\widehat{\nu}_q(\widehat{t})}{\|x\|_q^q} - 1\Big) \xrightarrow{w} N(0, v_q(c_0, \bar{\rho}_q)) \tag{3.31}$$

as $(n_q, p) \to \infty$, where the limiting variance is strictly positive and defined according to the formula

$$v_q(c_0, \bar{\rho}_q) := \tfrac{1}{|c_0|^{2q}}\Big(\tfrac{1}{2}\tfrac{1}{\varphi_0(\bar{\rho}_q|c_0|)^2}\exp(2|c_0|^q) + \tfrac{1}{2}\tfrac{\varphi_0(2\bar{\rho}_q|c_0|)}{\varphi_0(\bar{\rho}_q|c_0|)^2}\exp((2-2^q)|c_0|^q) - 1\Big). \tag{3.32}$$

Remarks. This result is proved in Appendix B.3. Although it might seem more natural to prove a CLT for the difference $\widehat{\nu}_q(\widehat{t}) - \|x\|_q^q$ rather than the ratio $\widehat{\nu}_q(\widehat{t})/\|x\|_q^q$, the advantage of the ratio is that its appropriate scaling factor is $\sqrt{n_q}$, and hence independent of the size of the (possibly growing) unknown parameter $\|x\|_q^q$.

Now that the limiting distribution of $\widehat{\nu}_q(\widehat{t})$ is available, we will focus on constructing an estimate $\widehat{t}$ so that the limiting value c_0 minimizes the variance function $v_q(\cdot, \bar{\rho}_q)$. Since the formula for $v_q(c_0, \bar{\rho}_q)$ is ill-defined for certain values of c_0, the following subsection extends the domain of $v_q(\cdot, \bar{\rho}_q)$ so that minimization can be formulated in a way that is more amenable to analysis.

Extending the variance function

Based on the previous theorem, our aim is to construct $\widehat{t}$ so that as $(n_q, p) \to \infty$,

$$\gamma_q \widehat{t} \|x\|_q \to_P c^\star(\bar{\rho}_q), \tag{3.33}$$

where $c^\star(\bar{\rho}_q)$ denotes a minimizer of $v_q(\cdot,\bar{\rho}_q)$. Since we assume the noise distribution is symmetric about 0, it follows that φ_0 is a symmetric function, and consequently $v_q(c,\rho)$ is symmetric in c. Therefore, for simplicity, we may restrict our attention to choices of c that are non-negative.

An inconvenient aspect of minimizing the function $v_q(\cdot,\rho)$ is that its domain depends on ρ — since the formula (3.32) is ill-defined at values of c_0 where $\varphi_0(\rho c_0)=0$. Because we will be interested in minimizing $v_q(\cdot,\widehat{\rho}_q)$ for some estimate $\widehat{\rho}_q$ of $\bar{\rho}_q$, this leads to analyzing the minimizer of a random function whose domain is also random. To alleviate this complication, we will define an extension of $v_q(\cdot,\cdot)$ whose domain does not depend on the second argument. Specifically, whenever $q\in(0,2)$, Proposition 3.2 below shows that an extension $\tilde{v}_q$ of v_q can be found with the properties that $\tilde{v}_q(\cdot,\cdot)$ is continuous on $[0,\infty)\times[0,\infty)$, and $\tilde{v}_q(\cdot,\bar{\rho}_q)$ has the same minimizers as $v_q(\cdot,\bar{\rho}_q)$.

When $q=2$, one additional detail must be handled. In this case, it may happen for certain noise distributions that $v_2(c,\bar{\rho}_2)$ approaches a minimum as c tends to 0 from the right[10], i.e.

$$\lim_{c\to 0+} v_2(c,\bar{\rho}_2)=\inf_{c>0} v_2(c,\bar{\rho}_2). \tag{3.34}$$

This creates a technical nuisance in using Theorem 3.2 because $v_2(c_0,\bar{\rho}_2)$ is not defined for $c_0=0$. There are various ways of handling this "edge case", but for simplicity, we take a practical approach of constructing $\widehat{t}$ so that

$$\gamma_2\widehat{t}\|x\|_2\to_P \varepsilon_2$$

for some (arbitrarily) small constant $\varepsilon_2>0$.[11] For this reason, in the particular case of $q=2$, we will restrict the domain of the extended function $\tilde{v}_2(\cdot,\cdot)$ to be $[\varepsilon_2,\infty)\times[0,\infty)$. The following lemma summarizes the properties of the extended variance function that will be needed later on.

Lemma 3.2 (Extended variance function)**.** *Suppose that φ_0 satisfies the assumptions of the model* (M)*, and let v_q be as in formula* (3.32)*. For each $q\in(0,2)$, put $\varepsilon_q:=0$, and let $\varepsilon_2>0$. For all values $q\in(0,2]$, define the function*

$$\tilde{v}_q:[\varepsilon_q,\infty)\times[0,\infty)\to[0,\infty], \tag{3.35}$$

according to

$$\tilde{v}_q(c,\rho):=\begin{cases} v_q(c,\rho) & \text{if } (c,\rho) \text{ satisfies } \ c\neq 0 \ \text{ and } \ \varphi_0(\rho c)\neq 0,\\ +\infty & \text{otherwise.}\end{cases} \tag{3.36}$$

Then, the function $\tilde{v}_q(\cdot,\cdot)$ is continuous on $[\varepsilon_q,\infty)\times[0,\infty)$, and for any $\rho\geq 0$, the function $\tilde{v}_q(\cdot,\rho)$ attains its minimum in the set $[\varepsilon_q,\infty)$.

[10] For instance, it can be checked that this occurs in the presence of noiseless measurements where $\varphi_0\equiv 1$, or when the noise distribution is Gaussian. However, for heavier tailed noise distributions, it can also happen that $v_2(\cdot,\bar{\rho}_q)$ is minimized at strictly positive values.

[11] An alternative solution is to simply avoid $q=2$ and estimate $s_q(x)$ for some q close to 2.

Remarks. A simple consequence of the definition of $\tilde{v}_q(\cdot,\cdot)$ is that any choice of $c \in [\varepsilon_q, \infty)$ that minimizes $\tilde{v}_q(\cdot, \bar{\rho}_q)$ also minimizes $v_q(\cdot, \bar{\rho}_q)$. Hence there is nothing lost in working with $\tilde{v}_q$.

Minimizing the extended variance function. We are now in position to specify the desired limiting value $c^\star(\bar{\rho}_q)$ from line (3.33). That is, for any $\rho \geq 0$, we define

$$c^\star(\rho) \in \underset{c \geq \varepsilon_q}{\operatorname{argmin}} \; \tilde{v}_q(c, \rho), \tag{3.37}$$

where $q \in (0, 2]$ and ε_q is as defined in Lemma 3.2. Note that $\tilde{v}_q$ is a known function, and so the value $c^\star(\rho)$ can be computed for any given ρ. However, since the limiting noise-to-signal ratio $\bar{\rho}_q$ is unknown, it will be necessary to work with $c^\star(\widehat{\rho}_q)$ for some estimate $\widehat{\rho}_q$ of $\bar{\rho}_q$, which is discussed below as part of our method for constructing an optimal $\widehat{t}$. Readers wishing to bypass the technical details of the procedure should skip ahead to Section 3.4.

A procedure for optimal selection of t

At a high level, we choose t by first computing a simple "pilot" value $\widehat{t}_{\text{pilot}}$, and then refining it to obtain an optimal value $\widehat{t}_{\text{opt}}$ that will be shown to satisfy

$$\widehat{t}_{\text{opt}} \gamma_q \|x\|_q \to_P c^\star(\bar{\rho}_q). \tag{3.38}$$

The pilot value of t will satisfy

$$\widehat{t}_{\text{pilot}} \gamma_q \|x\|_q \to_P c_0 \tag{3.39}$$

for some (possibly non-optimal) constant c_0. The construction of $\widehat{t}_{\text{pilot}}$ will be given in a moment. The purpose of the pilot value is to derive a ratio-consistent estimator of $\|x\|_q^q$ through the statistic $\widehat{\nu}_q(\widehat{t}_{\text{pilot}})$. With such an estimate of $\|x\|_q^q$ in hand, we can easily derive a consistent estimator of $\bar{\rho}_q$, namely

$$\widehat{\rho}_q := \tfrac{\sigma}{\gamma_q (\widehat{\nu}_q(\widehat{t}_{\text{pilot}}))^{1/q}}. \tag{3.40}$$

Next, we use $\widehat{\rho}_q$ to estimate the optimal constant $c^\star(\bar{\rho}_q)$ with $c^\star(\widehat{\rho}_q)$, as defined in line (3.37). Finally, we obtain an optimal choice of t using

$$\widehat{t}_{\text{opt}} := \tfrac{c^\star(\widehat{\rho}_q)}{\gamma_q (\widehat{\nu}_q(\widehat{t}_{\text{pilot}}))^{1/q}}. \tag{3.41}$$

The consistency of $\widehat{\rho}_q$ and $\widehat{t}_{\text{opt}}$ will be shown in Section 3.4.

Constructing the pilot value. In choosing a pilot value for t, there are two obstacles to consider. First, we must choose $\widehat{t}_{\text{pilot}}$ so that the limit (3.39) holds for some constant c_0. Second, we must ensure that the value c_0 is not a singularity of the function $v(\cdot, \bar{\rho}_q)$, i.e.

$c_0 \neq 0$ and $\varphi_0(\bar{\rho}_q c_0) \neq 0$, for otherwise the variance of $\widehat{\nu}_q(\widehat{t}_{\text{pilot}})$ may diverge as $(n_q, p) \to \infty$. To handle the first item, consider the median absolute deviation statistic

$$\widehat{m}_q := \text{med}(|y_1|, \dots, |y_{n_q}|), \tag{3.42}$$

which is a coarse-grained, yet robust, estimate of $\gamma_q \|x\|_q$. (The drawback of $\widehat{m}_q$ is that it does not deconvolve the effects of noise in estimating $\gamma_q \|x\|_q$.) If we define

$$\widehat{t}_{\text{initial}} := 1/\widehat{m}_q, \tag{3.43}$$

then a straightforward argument (see the proof of Proposition 3.3 in Section 3.4) shows there is a finite constant $c_1 > 0$ such that as $(n_q, p) \to \infty$,

$$\widehat{t}_{\text{initial}} \gamma_q \|x\|_q \to_P c_1. \tag{3.44}$$

Now, only a slight modification of $\widehat{t}_{\text{initial}}$ is needed so that the limiting constant c_1 avoids the singularities of $v_q(\cdot, \bar{\rho}_q)$. Since every characteristic function is continuous and satisfies $\varphi_0(0) = 1$, we may find a number $\eta_0 > 0$ such that $\varphi_0(\eta) > \frac{1}{2}$ for all $\eta \in [0, \eta_0]$. The value $\frac{1}{2}$ has no special importance. Using $\widehat{t}_{\text{initial}}$, we define

$$\widehat{t}_{\text{pilot}} := \widehat{t}_{\text{initial}} \wedge \tfrac{\eta_0}{\sigma},$$

where $a \wedge b = \min\{a, b\}$. Combining the limit (3.44) with assumption **A3.3**, it follows that $\widehat{t}_{\text{pilot}} \gamma_q \|x\|_q \to_P c_0$ for some finite constant $c_0 > 0$, since

$$\begin{aligned} \widehat{t}_{\text{pilot}} \gamma_q \|x\|_q &= (\widehat{t}_{\text{initial}} \gamma_q \|x\|_q) \wedge \left(\eta_0 \tfrac{\gamma_q \|x\|_q}{\sigma}\right) && (3.45) \\ &= c_1 \wedge \left(\tfrac{\eta_0}{\bar{\rho}_q}\right) + o_P(1) && (3.46) \\ &=: c_0 + o_P(1). && (3.47) \end{aligned}$$

Furthermore, it is clear that c_0 is not a singularity of $v(\cdot, \bar{\rho}_q)$, since c_0 is positive, and

$$\varphi_0(\bar{\rho}_q c_0) = \varphi_0\big((\bar{\rho}_q c_1) \wedge \eta_0\big) > \tfrac{1}{2},$$

due to the choice of η_0. This completes the description of $\widehat{t}_{\text{pilot}}$.

Algorithm for estimating $\|x\|_q^q$ and $s_q(x)$

We now summarize our method by giving a line-by-line algorithm for computing the adaptive estimator $\widehat{\nu}_q(\widehat{t}_{\text{opt}})$ of the parameter $\|x\|_q^q$. As described earlier, an estimate for $s_q(x)$ is obtained by combining norm estimates $\widehat{\nu}_1$ and $\widehat{\nu}_q$. When estimating $s_q(x)$ for $q \in (0, 2]$ and $q \neq 1$, we assume that two sets of measurements (of sizes n_1 and n_q) from the model (M) are available, i.e.

$$y_i = \langle a_i, x \rangle + \sigma \epsilon_i, \quad \text{with} \quad a_i \overset{\text{i.i.d.}}{\sim} \text{stable}_q(\gamma_q)^{\otimes p}, \quad \text{for} \quad i = 1, \dots, n_q, \tag{3.48}$$

$$y_i = \langle a_i, x \rangle + \sigma \epsilon_i, \quad \text{with} \quad a_i \overset{\text{i.i.d.}}{\sim} \text{stable}_1(\gamma_1)^{\otimes p}, \quad \text{for} \quad i = n_q + 1, \dots, n_q + n_1. \tag{3.49}$$

Once the estimators $\widehat{\nu}_q(\widehat{t}_{\text{opt}})$ and $\widehat{\nu}_1(\widehat{t}_{\text{opt}})$ have been computed with their respective values of $\widehat{t}_{\text{opt}}$, the estimator $\widehat{s}_q(x)$ is obtained as

$$\widehat{s}_q(x) := \frac{\left(\widehat{\nu}_q(\widehat{t}_{\text{opt}})\right)^{\frac{1}{1-q}}}{\left(\widehat{\nu}_1(\widehat{t}_{\text{opt}})\right)^{\frac{q}{1-q}}}. \tag{3.50}$$

The algorithm to compute $\widehat{\nu}_q(\widehat{t}_{\text{opt}})$ is given below.

Algorithm (Estimation procedure for $\|x\|_q^q$, for $q \in (0, 2]$).

```
Input: • observations $y_i$ generated with i.i.d. measurement vectors $a_i \sim \text{stable}_q(\gamma_q)^{\otimes p}$, for $i = 1, \ldots, n_q$
       • measurement intensity $\gamma_q$
       • noise level $\sigma$
       • noise characteristic function $\varphi_0$
       • threshold $\varepsilon_q$ defined in Lemma 3.2

1. compute $\widehat{t}_{\text{initial}} := 1/\widehat{m}_q$ where $\widehat{m}_q := \text{med}(|y_1|, \ldots, |y_{n_q}|)$
2. find $\eta_0 > 0$ such that $\varphi_0(\eta) > \frac{1}{2}$ for all $\eta \in [0, \eta_0]$
3. compute $\widehat{t}_{\text{pilot}} := \widehat{t}_{\text{initial}} \wedge \frac{\eta_0}{\sigma}$,
4. compute $\widehat{\rho}_q := \frac{\sigma}{\gamma_q(\widehat{\nu}_q(\widehat{t}_{\text{pilot}}))^{1/q}}$
5. compute $c^\star(\widehat{\rho}_q) \in \text{argmin}_{c \geq \varepsilon_q} \tilde{v}_q(c, \widehat{\rho}_q)$.
6. compute $\widehat{t}_{\text{opt}} := \frac{c^\star(\widehat{\rho}_q)}{\gamma_q(\widehat{\nu}_q(\widehat{t}_{\text{pilot}}))^{1/q}}$
7. return $\widehat{\nu}_q(\widehat{t}_{\text{opt}})$
```

3.4 Main results for estimators and confidence intervals

In this section, we first show in Proposition 3.3 that the procedure for selecting the $\widehat{t}_{\text{pilot}}$ leads to consistent estimates of the parameters $\|x\|_q^q$, and $\bar{\rho}_q$. Next, we show that the optimal constant $c^\star(\bar{\rho}_q)$ and optimal variance $v(c^\star(\bar{\rho}_q), \bar{\rho}_q)$ can also be consistently estimated. These estimators then lead to adaptive confidences intervals for $\|x\|_q^q$ and $s_q(x)$ based upon $\widehat{\nu}_q(\widehat{t}_{\text{opt}})$ in Theorem 3.3 and Corollary 3.1.

Consistency results

Proposition 3.3 (Consistency of the pilot estimator)**.** *Let $q \in (0,2]$ and assume that the measurement model* (M) *and* ***A3.3*** *hold. Then as $(n_q, p) \to \infty$,*

$$\frac{\widehat{\nu}_q(\widehat{t}_{\text{pilot}})}{\|x\|_q^q} \longrightarrow_P 1 \tag{3.51}$$

and

$$\widehat{\rho}_q \longrightarrow_P \bar{\rho}_q. \tag{3.52}$$

Remarks. We now turn our attention from the pilot value $\widehat{t}_{\text{pilot}}$ to the optimal value $\widehat{t}_{\text{opt}}$. In order to construct $\widehat{t}_{\text{opt}}$, our method relies on the M-estimator $c^\star(\widehat{\rho}_q) \in \operatorname{argmin}_{c \geq \varepsilon_q} v_q(c, \widehat{\rho}_q)$, where $c^\star(\cdot)$ is defined in line (3.37). Consistency proofs for M-estimators typically require that the objective function has a unique optimizer, and our situation is no exception. Note that we do not need to assume that a minimizer exists, since this is guaranteed by Proposition 3.2.

A3.4. *The function $v_q(\cdot, \bar{\rho}_q)$ has at most one minimizer in $[\varepsilon_q, \infty)$, where ε_q is defined in Proposition 3.2.*

In an approximate sense, this assumption can be verified empirically by simply plotting the function $v_q(\cdot, \widehat{\rho}_q)$. In Section B.6 of Appendix B, we verify the assumption analytically in the case of stable$_q$ noise. Based on graphical inspection, the assumption also seems to hold for a variety of natural parametric noise distributions (e.g. Laplace, uniform$[-1,1]$, and the t distribution). However, outside of special cases, analytic verification seems to be difficult, and even the stable case is somewhat involved.

Proposition 3.4 (Consistency of $c^\star(\widehat{\rho}_q)$)**.** *Let $q \in (0,2]$ and assume that the measurement model* (M) *holds, as well as assumptions* ***A3.3****, and* ***A3.4****. Then, as $(n_q, p) \to \infty$,*

$$c^\star(\widehat{\rho}_q) \ \longrightarrow_P \ c^\star(\bar{\rho}_q). \tag{3.53}$$

Furthermore,

$$\widehat{t}_{\text{opt}} \gamma_q \|x\|_q \ \longrightarrow_P \ c^\star(\bar{\rho}_q), \tag{3.54}$$

and

$$v_q(c^\star(\widehat{\rho}_q), \widehat{\rho}_q) \ \longrightarrow_P \ v_q(c^\star(\bar{\rho}_q), \bar{\rho}_q). \tag{3.55}$$

Remarks. This result is proved in Section B.4 of Appendix B.

Confidence intervals for $\|x\|_q^q$ and $s_q(x)$

In this subsection, we assemble the work in Proposition 3.4 with Theorem 3.2 to obtain confidence intervals for $\|x\|_q^q$ and $s_q(x)$. In the next two results, we will use $z_{1-\alpha}$ to denote the $1-\alpha$ quantile of the standard normal distribution, i.e. $\Phi(z_{1-\alpha}) = 1-\alpha$. To allow our result to be applied to both one-sided and two-sided intervals, we state our result in terms of two possibly distinct quantiles $z_{1-\alpha}$ and $z_{1-\alpha'}$.

Theorem 3.3 (Confidence interval for $\|x\|_q^q$)**.** *Let $q \in (0,2]$ and define the estimated variance*

$$\widehat{\omega}_q := v_q(c^\star(\widehat{\rho}_q), \widehat{\rho}_q). \tag{3.56}$$

Assume that the measurement model (M) *holds, as well as assumptions **A3.3**, and **A3.4**. Then as $(n_q, p) \to \infty$,*

$$\tfrac{\sqrt{n_q}}{\sqrt{\widehat{\omega}_q}}\Big(\tfrac{\widehat{\nu}_q(\widehat{t}_{\text{opt}})}{\|x\|_q^q} - 1\Big) \xrightarrow{w} N(0,1), \tag{3.57}$$

and consequently for any fixed $\alpha, \alpha' \in [0, \frac{1}{2}]$,

$$\mathbb{P}\left[\Big(1 - \tfrac{\sqrt{\widehat{\omega}_q} z_{1-\alpha}}{\sqrt{n_q}}\Big) \cdot \widehat{\nu}_q(\widehat{t}_{\text{opt}}) \leq \ \|x\|_q^q \ \leq \Big(1 + \tfrac{\sqrt{\widehat{\omega}_q} z_{1-\alpha'}}{\sqrt{n_q}}\Big) \cdot \widehat{\nu}_q(\widehat{t}_{\text{opt}})\right] \to 1 - \alpha - \alpha'. \tag{3.58}$$

Remarks. This result follows by combining Theorem 3.2 with Proposition 3.4. However, if the limit (3.57) is used directly to obtain a confidence interval for $\|x\|_q^q$, the resulting formulas are somewhat cumbersome. Instead, a simpler confidence interval (given in line (3.58)) is obtained using a CLT for the reciprocal $\|x\|_q^q / \widehat{\nu}_q(\widehat{t}_{\text{opt}})$, via the delta method. For a one-sided interval with the right endpoint being $+\infty$, we set $\alpha' = 0$, and similarly, we set $\alpha = 0$ in the opposite case.

As a corollary of Theorem 3.3, we obtain a CLT and a confidence interval for $\widehat{s}_q(x)$ by combining the estimators $\widehat{\nu}_q(\widehat{t}_{\text{opt}})$ and $\widehat{\nu}_1(\widehat{t}_{\text{opt}})$. Since each of the norm estimators rely on measurement sets of sizes n_q and n_1, we make the following simple scaling assumption, which enforces the idea that each set should be non-negligible with respect to the other.

A3.5. *For each $q \in (0,2] \setminus \{1\}$, there is a constant $\bar{\pi}_q \in (0,1)$, such that as $(n_1, n_q, p) \to \infty$,*

$$\tfrac{n_q}{n_1 + n_q} = \bar{\pi}_q + o(n_q^{-1/2}). \tag{3.59}$$

Corollary 3.1 (Confidence interval for $s_q(x)$)**.** *Assume $q \in (0,2] \setminus \{1\}$, and that the conditions of Theorem 3.3 hold, as well as assumption **A3.5**. Also assume $\widehat{s}_q(x)$ is constructed from independent sets of measurements* (3.48) *and* (3.49)*. Letting $\widehat{\omega}_q$ be as in Theorem 3.3, define the quantities*

$$\pi_q := n_q/(n_1 + n_q) \qquad \textit{and} \qquad \widehat{\vartheta}_q := \tfrac{\widehat{\omega}_q}{\pi_q}(\tfrac{1}{1-q})^2 + \tfrac{\widehat{\omega}_1}{1-\pi_q}(\tfrac{q}{1-q})^2.$$

Then as $(n_1, n_q, p) \to \infty$,

$$\tfrac{\sqrt{n_1+n_q}}{\sqrt{\widehat{\vartheta}_q}}\Big(\tfrac{\widehat{s}_q(x)}{s_q(x)} - 1\Big) \xrightarrow{w} N(0,1), \tag{3.60}$$

and consequently for any fixed $\alpha, \alpha' \in [0, \frac{1}{2}]$,

$$\mathbb{P}\Bigg[\Big(1 - \tfrac{\sqrt{\widehat{\vartheta}_q} z_{1-\alpha}}{\sqrt{n_1+n_q}}\Big) \cdot \widehat{s}_q(x) \leq \ s_q(x) \ \leq \Big(1 + \tfrac{\sqrt{\widehat{\vartheta}_q} z_{1-\alpha'}}{\sqrt{n_1+n_q}}\Big) \cdot \widehat{s}_q(x)\Bigg] \to 1 - \alpha - \alpha'. \tag{3.61}$$

Remarks. As in Theorem 3.3, we chose to present a simpler formula for the confidence interval in line (3.61) by using a CLT for the reciprocal $s_q(x)/\widehat{s}_q(x)$.

3.5 Applications of confidence intervals for $\|x\|_q^q$ and $s_q(x)$

In this section, we give some illustrative applications of our results for $\widehat{\|x\|_q^q}$ and $\widehat{s}_q(x)$. First, we describe how the assumption of sparsity may be checked in a hypothesis testing framework. Second, we consider the problem of choosing the regularization parameter for the Lasso and Elastic Net algorithms (in primal form).

Testing the hypothesis of sparsity

In the context of hypothesis testing, the null hypothesis is typically viewed as a "straw man" that the practitioner would like to reject in favor of the "more desirable" alternative hypothesis. Hence, for the purpose of verifying the assumption of sparsity, it is natural for the null hypothesis to correspond to a non-sparse signal. More specifically, if $1 < \kappa \leq p$ is a given reference value of sparsity, then we consider the testing problem

$$\mathbf{H}_0 : s_q(x) \geq \kappa \qquad \text{versus} \qquad \mathbf{H}_1 : 1 \leq s_q(x) < \kappa. \tag{3.62}$$

To construct a test statistic, we use the well-known duality between confidence intervals and hypothesis tests [LR05]. Consider a one-sided confidence interval for $s_q(x)$ of the form $(-\infty, \widehat{u}_\alpha]$, with asymptotic coverage probability $\mathbb{P}(s_q(x) \leq \widehat{u}_\alpha) = 1 - \alpha + o(1)$. Clearly, if $\mathbf{H}_0$ holds, then this one-sided interval must also contain κ with probability at least $1 - \alpha + o(1)$. Said differently, this means that under $\mathbf{H}_0$, the chance that $(-\infty, \widehat{u}_\alpha]$ fails to contain κ is at most $\alpha + o(1)$. Likewise, one may consider the test statistic

$$T := 1\big\{\widehat{u}_\alpha < \kappa\big\},$$

and reject $\mathbf{H}_0$ iff $T = 1$, which gives an asymptotically valid level-α testing procedure. Now, by Corollary 3.1, if we choose

$$\widehat{u}_\alpha := (1 + \tfrac{\widehat{\vartheta}_q z_{1-\alpha}}{\sqrt{n_1+n_q}})\widehat{s}_q(x),$$

then the interval $(-\infty, \widehat{u}_\alpha]$ has asymptotic coverage probability $1-\alpha$. The reasoning just given ensures that the false alarm rate is asymptotically bounded by α as $(n_1, n_q, p) \to \infty$. Namely,

$$\mathbb{P}_{\mathbf{H}_0}(T=1) \leq \alpha + o(1).$$

It is also possible to derive the asymptotic power function of the test statistic. Let $\widehat{\vartheta}_q$ be as defined in Corollary 3.1, and note that this variable converges in probability to a positive constant, say ϑ_q (by Proposition 3.4). Then, as $(n_1, n_q, p) \to \infty$, the asymptotic power satisfies

$$\mathbb{P}_{\mathbf{H}_1}(T=1) = \Phi\Big(\tfrac{\sqrt{n_1+n_q}}{\vartheta_q}\Big(\tfrac{\kappa}{s_q(x)} - 1\Big) - z_{1-\alpha}\Big) + o(1). \tag{3.63}$$

The details of obtaining this limit are straightforward, and hence omitted. Note that as $s_q(x)$ becomes close to the reference value κ (i.e. the detection boundary), the power approaches that of random guessing, $\Phi(-z_{1-\alpha}) = \alpha$, as we would expect.

Tuning the Lasso and Elastic Net in primal form

In primal form, the Lasso algorithm can be expressed as

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\text{minimize}} \quad & \|y - Av\|_2^2 \\ \text{subject to} \quad & v \in \mathbb{B}_1(r), \end{aligned} \tag{3.64}$$

where $\mathbb{B}_1(r) := \{v \in \mathbb{R}^p : \|v\|_1 \leq r\}$ is the ℓ_1 ball of radius $r \geq 0$, and r is viewed as the regularization parameter. (Note that the matrix A here may be different from the measurement matrix we use to estimate $s_q(x)$.) If x denotes the true signal, then $\mathbb{B}_1(\|x\|_1)$ is the smallest such set for which the true signal is feasible. Hence, one would expect $r = \|x\|_1$ to be an ideal choice of the tuning parameter. In fact, the recent paper [Cha14] shows that this intuition is correct in a precise sense by quantifying how the mean-squared prediction error of the Lasso deteriorates when the tuning parameter differs from $\|x\|_1$.

When using a data-dependent tuning parameter $\widehat{r}$, it is of interest to have some guarantee that the true signal is likely to lie in the (random) set $\mathbb{B}_1(\widehat{r})$. Our one-sided confidence interval for $\|x\|_1$ precisely solves this problem. More specifically, under the assumptions of Theorem 3.3, if we choose $\alpha = 0$, and $\alpha' \in [0, \frac{1}{2}]$, then under the choice

$$\widehat{r} := \Big(1 + \tfrac{\sqrt{\widehat{\omega}_1} z_{1-\alpha'}}{\sqrt{n_1}}\Big) \cdot \widehat{\nu}_1(\widehat{t}_{\text{opt}}), \tag{3.65}$$

we have as $(n_1, p) \to \infty$

$$\mathbb{P}\big(x \in \mathbb{B}_1(\widehat{r})\big) \to 1 - \alpha'. \tag{3.66}$$

In fact, this idea can be extended further by adding extra ℓ_q norm constraints. A natural example is a primal form of the well known *Elastic Net* algorithm [ZH05], which constrains both the ℓ_1 and ℓ_2 norms, leading to the convex program

$$\begin{aligned} \underset{v\in\mathbb{R}^p}{\text{minimize}} \quad & \|y - Av\|_2^2 \\ \text{subject to} \quad & v \in \mathbb{B}_1(r) \\ & v \in \mathbb{B}_2(\varrho) \end{aligned} \tag{3.67}$$

for some parameters $r, \varrho \geq 0$. Here, $\mathbb{B}_2$ is defined in the same way as $\mathbb{B}_1$. Again, under the assumptions of Theorem 3.3, if for some $\alpha \in [0, \frac{1}{2}]$ we put

$$\widehat{\varrho} := \Big(1 + \tfrac{\sqrt{\widehat{\omega}_2} z_{1-\alpha}}{\sqrt{n_2}}\Big) \cdot \widehat{\nu}_2(\widehat{t}_{\text{opt}}), \tag{3.68}$$

and if the respective measurement sets of size n_1 and n_2 used to construct $\widehat{r}$ and $\widehat{\varrho}$ are independent, then as $(n_1, n_2, p) \to \infty$,

$$\mathbb{P}\Big(x \in \mathbb{B}_1(\widehat{r}) \cap \mathbb{B}_2(\widehat{\varrho})\Big) \to (1-\alpha)^2. \tag{3.69}$$

The same reasoning applies to any other combination of ℓ_q norms for $q \in (0, 2]$.

3.6 Deterministic measurement matrices

The problem of constructing deterministic matrices A with good recovery properties (e.g. RIP-k or NSP-k) has been a longstanding open direction within CS [FR13, see Sections 1.3 and 6.1] [DeV07]. Since our procedure in Section 3.3 selects A at random, it is natural to ask if randomization is essential to the estimation of unknown sparsity. In this section, we show that estimating $s_q(x)$ with a deterministic matrix A leads to results that are inherently different from our randomized procedure.

At an informal level, the difference between random and deterministic matrices makes sense if we think of the estimation problem as a game between nature and a statistician. Namely, the statistician first chooses a matrix $A \in \mathbb{R}^{n\times p}$ and an estimation rule $\delta : \mathbb{R}^n \to \mathbb{R}$. (The function δ takes $y \in \mathbb{R}^n$ as input and returns an estimate of $s_q(x)$.) In turn, nature chooses a signal $x \in \mathbb{R}^p$, with the goal of maximizing the statistician's error. When the statistician chooses A deterministically, nature has the freedom to adversarially select an x that is ill-suited to the fixed matrix A. By contrast, if the statistician draws A at random, then nature does not know what value A will take, and therefore has less knowledge to choose a "bad" signal.

In the case of *random* measurements, Corollary 3.1 implies that our particular estimation rule $\widehat{s}_q(x)$ can achieve a relative error $|\widehat{s}_q(x)/s_q(x) - 1|$ on the order of $1/\sqrt{n_1 + n_q}$ with high probability for any non-zero x. Our aim is now to show that for any set of noiseless *deterministic* measurements, *all* estimation rules $\delta : \mathbb{R}^n \to \mathbb{R}$ have a worst-case relative error $|\delta(Ax)/s_2(x) - 1|$ that is much larger than $1/\sqrt{n_1 + n_q}$. Specifically, when $q \in [0, 2]$, we give a lower bound that is of order $(1 - \frac{n}{p})^2$, which means that in the worst case, $s_q(x)$ cannot be estimated consistently in relative error when $n \ll p$. (The same conclusion holds for $q \in (2,\infty]$ up to a factor of $\sqrt{\log(2p)}$.) More informally, this means that there is always a choice of x that can defeat a deterministic procedure, whereas the randomized estimator $\widehat{s}_q(x)$ is likely to succeed under any choice of x.

In stating the following result, we note that it involves no randomness whatsoever — since we assume here that the observed measurements $y = Ax$ are noiseless and obtained from a deterministic matrix A. Furthermore, the bounds are non-asymptotic.

Theorem 3.4. *Suppose $n < p$, and $q \in [0,2]$. Then, the minimax relative error for estimating $s_q(x)$ from noiseless deterministic measurements $y = Ax$ satisfies*

$$\inf_{A\in\mathbb{R}^{n\times p}} \inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\mathbb{R}^p\setminus\{0\}} \left|\tfrac{\delta(Ax)}{s_q(x)} - 1\right| \geq \tfrac{1}{2\pi e}\left(1 - \tfrac{n}{p}\right)^2 - \tfrac{1}{2p}.$$

Alternatively, if $q \in (2,\infty]$, then

$$\inf_{A\in\mathbb{R}^{n\times p}} \inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\mathbb{R}^p\setminus\{0\}} \left|\tfrac{\delta(Ax)}{s_q(x)} - 1\right| \geq \tfrac{1}{\sqrt{2\pi e}} \cdot \tfrac{1-(n/p)}{1+\sqrt{16\log(2p)}} - \tfrac{1}{2p}.$$

Remarks. The proof of this result is based on the classical technique of a *two-point prior.* In essence, the idea is that for any choice of A, it is possible to find two signals $\tilde{x}$ and x° that are indistinguishable with respect to A, i.e.

$$A\tilde{x} = Ax^\circ, \tag{3.70}$$

and yet have very different sparsity levels,

$$s_q(x^\circ) \ll s_q(\tilde{x}). \tag{3.71}$$

Due to the relation (3.70), the statistician has no way of knowing whether $\tilde{x}$ or x° has been selected by nature, and if nature chooses x° and $\tilde{x}$ with equal probability, then it is impossible for the statistician to improve upon the trivial estimator $\frac{1}{2}s_q(\tilde{x}) + \frac{1}{2}s_q(x^\circ)$ that does not even make use of the data. Furthermore, since $s_q(x^\circ) \ll s_q(\tilde{x})$, it follows that the trivial estimator has a large relative error – implying that the minimax relative error is also large. (A formal version of this argument is given in Section B.5 of Appendix B.)

To implement the approach of a two-point prior, the main challenge is to show that for *any* choice of $A \in \mathbb{R}^{n\times p}$, two vectors satisfying (3.70) and (3.71) can actually be found. This is the content of the following lemma.

Lemma 3.3. *Let $A \in \mathbb{R}^{n\times p}$ be an arbitrary matrix with $n < p$, and let $x^\circ \in \mathbb{R}^p$ be an arbitrary signal. Then, for each $q \in [0,2]$, there exists a non-zero vector $\tilde{x} \in \mathbb{R}^p$ satisfying $A\tilde{x} = Ax^\circ$ and*

$$s_q(\tilde{x}) \geq \tfrac{1}{\pi e} \cdot (1 - \tfrac{n}{p})^2 \cdot p. \tag{3.72}$$

Also, for $q \in (2,\infty]$, there is a vector $\bar{x}$ satisfying $A\bar{x} = Ax^\circ$ and

$$s_q(\bar{x}) \geq \frac{\sqrt{\frac{2}{\pi e}}(p-n)}{1 + \sqrt{16\log(2p)}}. \tag{3.73}$$

Remarks. Although it might seem intuitively obvious that every affine subspace contains a non-sparse vector, the technical substance of the result lies in the fact that the bounds hold *uniformly* over all matrices $A \in \mathbb{R}^{n\times p}$. This uniformity is necessary when taking the infimum over all $A \in \mathbb{R}^{n\times p}$ in Theorem 3.4. Furthermore, the order of magnitude of the bounds for $q \in [0, 2]$ is unimprovable when $n \ll p$, since $s_q(x) \leq p$. Similarly, the bound for $q \in (2, \infty]$ is optimal up to a logarithmic factor. Our proof in Appendix B.5 uses the probabilistic method to show that the desired vectors $\tilde{x}$ and $\bar{x}$ exist. Namely, we put a distribution on the set of vectors v satisfying $Ax = Av$, and then show that the stated bounds hold with positive probability.

Bibliography

[Aiv+89] Aivazian, S. A. et al. “Applied Statistics (in Russian). vol. 3: Classification and Reduction of Dimension”. In: *Finansy i Statistika, Moscow* (1989).

[AMS96] Alon, N., Matias, Y., and Szegedy, M. “The space complexity of approximating the frequency moments”. In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing.* ACM. 1996, pp. 20–29.

[B+05] Butucea, C., Matias, C., et al. “Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model”. In: *Bernoulli* 11.2 (2005), pp. 309–340.

[BD01] Bickel, P. J. and Doksum, K. *Mathematical Statistics, volume I.* Prentice Hall, 2001.

[BDB07] Boufounos, P., Duarte, M. F., and Baraniuk, R. G. “Sparse signal reconstruction from noisy compressive measurements using cross validation”. In: *Statistical Signal Processing, 2007. SSP’07. IEEE/SP 14th Workshop on.* IEEE. 2007, pp. 299–303.

[Bel+61] Bellman, R. et al. *Adaptive control processes: a guided tour.* Vol. 4. Princeton university press Princeton, 1961.

[BF81] Bickel, P. J. and Freedman, D. A. “Some asymptotic theory for the bootstrap”. In: *The Annals of Statistics* (1981), pp. 1196–1217.

[BF83] Bickel, P. J. and Freedman, D. A. “Bootstrapping regression models with many parameters”. In: *Festschrift for Erich L. Lehmann.* Wadsworth, 1983, pp. 28–48.

[BG11] Bühlmann, P. and Geer, S. van de. *Statistics for high-dimensional data: methods, theory and applications.* Springer, 2011.

[BGZ97] Bickel, P. J., Gotze, F., and Zwet, W. van. “Resampling fewer than n observations: Gains, losses, and remedies for the losses”. In: *Statistica Sinica* 7 (1997), pp. 1–31.

[BH95] Benjamini, Y. and Hochberg, Y. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society. Series B* (1995), pp. 289–300.

[Bha97] Bhatia, R. *Matrix analysis.* Vol. 169. Springer, 1997.

[Bic+09] Bickel, P. J. et al. "An overview of recent developments in genomics and associated statistical methods". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367.1906 (2009), pp. 4313–4337.

[BKS14] Barak, B., Kelner, J. A., and Steurer, D. "Rounding sum-of-squares relaxations". In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM. 2014, pp. 31–40.

[BL00] Borwein, J. M. and Lewis, A. S. *Convex Analysis and Nonlinear Optimization Theory and Examples*. CMS Books in Mathematics. Canadian Mathematical Society, 2000.

[BL08] Bickel, P. J. and Levina, E. "Covariance regularization by thresholding". In: *The Annals of Statistics* (2008), pp. 2577–2604.

[BL14] Bobkov, S. and Ledoux, M. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*. preprint, 2014.

[BLM13] Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BS10] Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*. Vol. 20. Springer, 2010.

[BS96] Bai, Z. and Saranadasa, H. "Effect of high dimension: by an example of a two sample problem". In: *Statistica Sinica* 6 (1996), pp. 311, 329.

[Buh13] Bühlmann, P. "Statistical significance in high-dimensional linear models". In: *Bernoulli* 19.4 (2013), pp. 1212–1242.

[BV04] Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004.

[Can06] Candès, E. "Compressive sampling". In: *Proceedings of the International Congress of Mathematicians*. 2006, pp. 1433–1452.

[CC12] Clifford, P. and Cosma, I. A. "A statistical analysis of probabilistic counting algorithms". In: *Scandinavian Journal of Statistics* 39.1 (2012), pp. 1–14.

[CDD09] Cohen, A., Dahmen, W., and DeVore, R. "Compressed sensing and best k-term approximation". In: *Jouran of the American Mathematical Society* 22.1 (2009), pp. 211–231.

[CDS98] Chen, S. S., Donoho, D. L., and Saunders, M. A. "Atomic decomposition by basis pursuit". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 33–61.

[Cha+08] Chan, W. L. et al. "Terahertz imaging with compressed sensing and phase retrieval". In: *Optics letters* 33.9 (2008), pp. 974–976.

[Cha+12] Chandrasekaran, V. et al. "The convex geometry of linear inverse problems". In: *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849.

[Cha14] Chatterjee, S. "A new perspective on least squares under convex constraint". In: *The Annals of Statistics* 42.6 (Dec. 2014), pp. 2340–2381.

[Chi03] Chikuse, Y. *Statistics on Special Manifolds*. Vol. 174. Springer, 2003.

[CHJ10] Calderbank, R., Howard, S., and Jafarpour, S. "Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property". In: *IEEE Journal of Selected Topics in Signal Processing* 4.2 (2010), pp. 358–374.

[CJ13] Chandrasekaran, V. and Jordan, M. I. "Computational and statistical tradeoffs via convex relaxation". In: *Proceedings of the National Academy of Sciences* 110.13 (2013), E1181–E1190.

[CL13] Chatterjee, A. and Lahiri, S. N. "Rates of convergence of the Adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap". In: *The Annals of Statistics* 41.3 (2013), pp. 1232–1259.

[Cor+03] Cormode, G. et al. "Comparing data streams using hamming norms (how to zero in)". In: *IEEE Transactions on Knowledge and Data Engineering* 15.3 (2003), pp. 529–540.

[Cor+12] Cormode, G. et al. "Synopses for massive data: Samples, histograms, wavelets, sketches". In: *Foundations and Trends in Databases* 4.1–3 (2012), pp. 1–294.

[Cou13] Council, N. R. *Frontiers in Massive Data Analysis*. The National Academies Press, 2013.

[CP11] Candès, E. J. and Plan, Y. "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements". In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2342–2359.

[CR09] Candès, E. and Recht, B. "Exact matrix completion via convex optimization". In: *Foundations of Computational Mathematics* 9.6 (2009), pp. 717–772.

[CRT06] Candès, E., Romberg, J., and Tao, T. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on pure and applied mathematics* 59.8 (2006), pp. 1207–1223.

[Cso81a] Csörgő, S. "Limit behaviour of the empirical characteristic function". In: *The Annals of Probability* (1981), pp. 130–144.

[Cso81b] Csörgő, S. "Multivariate empirical characteristic functions". In: *Probability Theory and Related Fields* 55.2 (1981), pp. 203–229.

[CT05] Candès, E. and Tao, T. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[CT07] Candès, E. and Tao, T. "The Dantzig selector: statistical estimation when p is much larger than n". In: *The Annals of Statistics* 35.6 (2007), pp. 2313–2351.

[CW92] Coifman, R. R. and Wickerhauser, M. V. "Entropy-based algorithms for best basis selection". In: *IEEE Transactions on Information Theory* 38.2 (1992), pp. 713–718.

[CWX10] Cai, T., Wang, L., and Xu, G. "New bounds for restricted isometry constants". In: *IEEE Transactions on Information Theory* 56.9 (2010).

[DE11] Duarte, M. F. and Eldar, Y. C. "Structured compressed sensing: From theory to applications". In: *IEEE Transactions on Signal Processing* 59.9 (2011), pp. 4053–4085.

[dE11] d'Aspremont, A. and El Ghaoui, L. "Testing the nullspace property using semidefinite programming". In: *Mathematical programming* 127.1 (2011), pp. 123–144.

[DeV07] DeVore, R. A. "Deterministic constructions of compressed sensing matrices". In: *Journal of Complexity* 23.4 (2007), pp. 918–925.

[DH01] Donoho, D. L. and Huo, X. "Uncertainty principles and ideal atomic decomposition". In: *IEEE Transactions on Information Theory* 47.7 (2001), pp. 2845–2862.

[DH14] Demanet, L. and Hand, P. "Scaling law for recovering the sparsest element in a subspace". In: *Information and Inference* (2014).

[DHG13] Deng, W., Hu, J., and Guo, J. "In defense of sparsity based face recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 399–406.

[DM09] Dai, W. and Milenkovic, O. "Subspace pursuit for compressive sensing signal reconstruction". In: *IEEE Transactions on Information Theory* 55.5 (2009), pp. 2230–2249.

[Don00] Donoho, D. L. "Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality". In: *American Mathematical Society Lecture: Math Challenges of the 21st Century* (2000).

[Don06] Donoho, D. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.

[Don94] Donoho, D. L. "On minimum entropy segmentation". In: *Wavelets: Theory, algorithms, and applications* (1994), pp. 233–269.

[DS98] Draper, N. R. and Smith, H. *Applied regression analysis*. Wiley-Interscience, 1998.

[Dua+08] Duarte, M. F. et al. "Single-pixel imaging via compressive sampling". In: *IEEE Signal Processing Magazine* 25.2 (2008), p. 83.

[Efr79] Efron, B. "Bootstrap methods: another look at the jackknife". In: *The Annals of Statistics* (1979), pp. 1–26.

[EK12] Eldar, Y. C. and Kutyniok, G. *Compressed Sensing: Theory and Applications.* Cambridge University Press, 2012.

[Eld09] Eldar, Y. "Generalized SURE for exponential families: Applications to regularization". In: *IEEE Transactions on Signal Processing* 57.2 (2009), pp. 471–481.

[EP15] El Karoui, N. and Purdom, E. "Can we trust the bootstrap in high-dimension?" In: *UC Berkeley Statistics Department Technical Report* (2015).

[ER05] Edelman, A. and Rao, N. R. "Random matrix theory". In: *Acta Numerica* 14 (2005), pp. 233–297.

[F+12] Feuerverger, A., He, Y., Khatri, S., et al. "Statistical significance of the Netflix challenge". In: *Statistical Science* 27.2 (2012), pp. 202–231.

[Fei+02] Feigenbaum, J. et al. "An approximate L 1-difference algorithm for massive data streams". In: *SIAM Journal on Computing* 32.1 (2002), pp. 131–151.

[Fou+10] Foucart, S. et al. "The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$". In: *Journal of Complexity* 26.6 (2010), pp. 629–640.

[Fou07] Foundation, N. S. *Discovery in Complex or Massive Datasets: Common Statistical Themes.* 2007.

[FR13] Foucart, S. and Rauhut, H. *A mathematical introduction to compressive sensing.* Springer, 2013.

[Fre81] Freedman, D. A. "Bootstrapping regression models". In: *The Annals of Statistics* 9.6 (1981), pp. 1218–1228.

[GG84] Garnaev, A. and Gluskin, E. "The widths of a Euclidean ball". In: *Dokl. Akad. Nauk SSSR.* Vol. 277. 5. 1984, pp. 1048–1052.

[GI10] Gilbert, A. and Indyk, P. "Sparse recovery using sparse matrices". In: *Proceedings of the IEEE.* Vol. 98. IEEE. 2010, pp. 937–947.

[Gne01] Gneiting, T. "Curiosities of characteristic functions". In: *Expositiones Mathematicae* 19.4 (2001), pp. 359–363.

[GS02] Gibbs, A. L. and Su, F. E. "On choosing and bounding probability metrics". In: *International Statistical Review* 70.3 (2002), pp. 419–435.

[Har69] Hartigan, J. A. "Using subsample values as typical values". In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1303–1317.

[Har75] Hartigan, J. A. "Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values". In: *The Annals of Statistics* 3.3 (1975), pp. 573–580.

[Hil73] Hill, M. O. "Diversity and evenness: a unifying notation and its consequences". In: *Ecology* 54.2 (1973), pp. 427–432.

[HJ09] Horn, R. A. and Johnson, C. A. *Matrix analysis.* 22nd printing. Cambridge University Press, 2009.

[HJ91] Horn, R. A. and Johnson, C. R. *Topics in matrix analysis.* 1991.

[Hoy04] Hoyer, P. O. "Non-negative matrix factorization with sparseness constraints". In: *The Journal of Machine Learning Research* 5 (2004), pp. 1457–1469.

[HR09] Hurley, N. and Rickard, S. "Comparing measures of sparsity". In: *IEEE Transactions on Information Theory* 55.10 (2009), pp. 4723–4741.

[HTF09] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Vol. 2. 1. Springer, 2009.

[Hub73] Huber, P. J. "Robust regression: asymptotics, conjectures and Monte Carlo". In: *The Annals of Statistics* (1973), pp. 799–821.

[Ili76] Il'inskii, A. "On the zeros and the argument of a characteristic function". In: *Theory of Probability & Its Applications* 20.2 (1976), pp. 410–415.

[Ind06] Indyk, P. "Stable distributions, pseudorandom generators, embeddings, and data stream computation". In: *Journal of the Association for Computing Machinery* 53.3 (2006), pp. 307–323.

[Ind13] Indyk, P. "Sketching via hashing: from heavy hitters to compressed sensing to sparse fourier transform". In: *Proceedings of the 32nd symposium on Principles of database systems.* ACM. 2013, pp. 87–90.

[Jam54] James, A. T. "Normal multivariate analysis and the orthogonal group". In: *The Annals of Mathematical Statistics* (1954), pp. 40–75.

[JM14a] Javanmard, A. and Montanari, A. "Hypothesis Testing in High-Dimensional Regression Under the Gaussian Random Design Model: Asymptotic Theory". In: *IEEE Transactions on Information Theory* 60.10 (Oct. 2014), pp. 6522–6554.

[JM14b] Javanmard, A. and Montanari, A. "Confidence intervals and hypothesis testing for high-dimensional regression". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2869–2909.

[JN11] Juditsky, A. and Nemirovski, A. "On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization". In: *Mathematical programming* 127.1 (2011), pp. 57–88.

[Joh13] Johnstone, I. M. *Gaussian estimation: Sequence and wavelet models Draft of a monograph.* 2013.

[Jos06] Jost, L. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[Kas77] Kashin, B. "Diameters of some finite-dimensional sets and classes of smooth functions". In: *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 41.2 (1977), pp. 334–351.

[KLT11] Koltchinskii, V., Lounici, K., and Tsybakov, A. B. "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion". In: *The Annals of Statistics* 39.5 (2011), pp. 2302–2329.

[Kol09] Kolaczyk, E. D. *Statistical Analysis of Network Data: Methods and Models.* Springer, 2009.

[Las09] Lasserre, J.-B. *Moments, positive polynomials and their applications.* Vol. 1. World Scientific, 2009.

[Led05] Ledoux, M. *The concentration of measure phenomenon.* 89. American Mathematical Soc., 2005.

[LHC07] Li, P., Hastie, T. J., and Church, K. W. "Nonlinear estimators and tail bounds for dimension reduction in l_1 using Cauchy random projections". In: *Learning Theory.* Springer, 2007, pp. 514–529.

[Li08] Li, P. "Estimators and tail bounds for dimension reduction in l_α $(0 < \alpha \leq 2)$ using stable random projections". In: *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms.* Society for Industrial and Applied Mathematics. 2008, pp. 10–19.

[LJW11] Lopes, M., Jacob, L., and Wainwright, M. "A More Powerful Two-Sample Test in High Dimensions using Random Projection". In: *Advances in Neural Information Processing Systems (NIPS).* 2011, pp. 1206–1214.

[LM00] Laurent, B. and Massart, P. "Adaptive Estimation of a Quadratic Functional by Model Selection". In: *Annals of Statistics* 28.5 (2000), pp. 1302–1338.

[Lop13] Lopes, M. "Estimating Unknown Sparsity in Compressed Sensing". In: *Proceedings of The 30th International Conference on Machine Learning (ICML).* 2013, pp. 217–225.

[Lop14] Lopes, M. "A Residual Bootstrap for High-Dimensional Regression with Near Low-Rank Designs". In: *Advances in Neural Information Processing Systems (NIPS).* 2014, pp. 3239–3247.

[LR05] Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses.* Springer, 2005.

[LR09] Lévy-Leduc, C. and Roueff, F. "Detection and localization of change-points in high-dimensional network traffic data". In: *The Annals of Applied Statistics* (2009), pp. 637–662.

[LY13] Liu, H. and Yu, B. "Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression". In: *Electronic Journal of Statistics* 7 (2013), pp. 3124–3169.

[LZZ14] Li, P., Zhang, C.-H., and Zhang, T. "Compressed Counting Meets Compressed Sensing". In: *Journal of Machine Learning Research, Workshop and Conference Proceedings.* 2014, pp. 1058–1077.

[Mac09] MacKenzie, D. "Compressed sensing makes every pixel count". In: *What's happening in the mathematical sciences* 7 (2009), pp. 114–127.

[Mah46] Mahalanobis, P. "Sample surveys of crop yields in India". In: *Sankhyā: The Indian Journal of Statistics* (1946), pp. 269–280.

[Mal08] Mallat, S. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[Mam89] Mammen, E. "Asymptotics with increasing dimension for robust regression with applications to the bootstrap". In: *The Annals of Statistics* (1989), pp. 382–400.

[Mar81] Marcus, M. B. "Weak convergence of the empirical characteristic function". In: *The Annals of Probability* 9.2 (1981), pp. 194–201.

[Mas00] Massart, P. "Some applications of concentration inequalities to statistics". In: *Annales-Faculte des Sciences Toulouse Mathematiques*. Vol. 9. 2. Université Paul Sabatier. 2000, pp. 245–303.

[Mat02] Matias, C. "Semiparametric deconvolution with unknown noise variance". In: *ESAIM: Probability and Statistics* 6 (2002), pp. 271–292.

[Mat97] Mathai, A. M. *Jacobians of matrix transformations and functions of matrix argument*. World Scientific, 1997.

[MB06] Meinshausen, N. and Bühlmann, P. "High-dimensional graphs and variable selection with the lasso". In: *The Annals of Statistics* (2006), pp. 1436–1462.

[Mei06] Meister, A. "Density estimation with normal measurement error with unknown variance". In: *Statistica Sinica* 16.1 (2006), p. 195.

[MH95] Markatou, M. and Horowitz, J. L. "Robust scale estimation in the error-components model using the empirical characteristic function". In: *Canadian Journal of Statistics* 23.4 (1995), pp. 369–381.

[MHL95] Markatou, M., Horowitz, J. L., and Lenth, R. V. "Robust scale estimation based on the the empirical characteristic function". In: *Statistics & probability letters* 25.2 (1995), pp. 185–192.

[MOA10] Marshall, A., Olkin, I., and Arnold, B. *Inequalities: theory of majorization and its applications*. Springer, 2010.

[MSW08] Malioutov, D., Sanghavi, S., and Willsky, A. "Compressed sensing with sequential observations". In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008*. IEEE. 2008, pp. 3357–3360.

[Mui82] Muirhead, R. J. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, 1982.

[MZ93] Mallat, S. G. and Zhang, Z. "Matching pursuits with time-frequency dictionaries". In: *IEEE Transactions on Signal Processing* 41.12 (1993), pp. 3397–3415.

[Neg+12] Negahban, S. N. et al. "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers". In: *Statistical Science* 27.4 (2012), pp. 538–557.

[NT09] Needell, D. and Tropp, J. A. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321.

[NW12] Negahban, S. and Wainwright, M. J. "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 1665–1697.

[Nyq28] Nyquist, H. "Certain topics in telegraph transmission theory". In: *Transactions of the American Institute of Electrical Engineers* 47.2 (1928), pp. 617–644.

[Pet94] Petz, D. "A survey of certain trace inequalities". In: *Functional analysis and operator theory.* Banach Center Publications 30 (1994), pp. 287–298.

[PGC12] Pilanci, M., Ghaoui, L. E., and Chandrasekaran, V. "Recovery of sparse probability measures via convex programming". In: *Advances in Neural Information Processing Systems (NIPS).* 2012, pp. 2420–2428.

[Por84] Portnoy, S. "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency". In: *The Annals of Statistics* (1984), pp. 1298–1309.

[PRK93] Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition". In: *Asilomar Conference on Signals, Systems and Computers.* IEEE. 1993, pp. 40–44.

[PRW99] Politis, D., Romano, J., and Wolf, M. *Subsampling.* Springer, 1999.

[Que49] Quenouille, M. H. "Approximate tests of correlation in time-series 3". In: *Mathematical Proceedings of the Cambridge Philosophical Society.* Vol. 45. 03. 1949, pp. 483–484.

[RBL11] Rigamonti, R., Brown, M., and Lepetit, V. "Are sparse representations really relevant for image classification?" In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2011, pp. 1545–1552.

[Rep+15] Repetti, A. et al. "Euclid in a Taxicab: Sparse Blind Deconvolution with Smoothed ℓ_1/ℓ_2 Regularization". In: *IEEE Signal Processing Letters* 22.5 (2015), pp. 539–543.

[RFP10] Recht, B., Fazel, M., and Parrilo, P. A. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[RK99] Rao, B. D. and Kreutz-Delgado, K. "An affine scaling methodology for best basis selection". In: *Signal Processing, IEEE Transactions on* 47.1 (1999), pp. 187–200.

[RV07] Rudelson, M. and Vershynin, R. "Sampling from large matrices: An approach through geometric functional analysis". In: *Journal of the Association for Computing Machinery* 54.4 (2007), p. 21.

[RV97] Resnick, P. and Varian, H. R. "Recommender systems". In: *Communications of the ACM* 40.3 (1997), pp. 56–58.

[RWY11] Raskutti, G., Wainwright, M. J., and Yu, B. "Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-Balls". In: *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6976–6994.

[Ser08] Serdobolskii, V. I. *Multiparametric statistics*. Elsevier, 2008.

[Sha49] Shannon, C. E. "Communication in the presence of noise". In: *Proceedings of the Institute of Radio Engineers* 37.1 (1949), pp. 10–21.

[Shi+11] Shi, Q. et al. "Is face recognition really a Compressive Sensing problem?" In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2011, pp. 553–560.

[SR05] Silverman, B. and Ramsay, J. *Functional data analysis*. Springer, 2005.

[ST95] Shao, J. and Tu, D. *The jackknife and bootstrap*. Springer, 1995.

[Sto74] Stone, M. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society. Series B* (1974), pp. 111–147.

[TG07] Tropp, J. A. and Gilbert, A. C. "Signal recovery from random measurements via orthogonal matching pursuit". In: *IEEE Transactions on Information Theory* 53.12 (2007), pp. 4655–4666.

[Tib96] Tibshirani, R. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[TN11] Tang, G. and Nehorai, A. "Performance analysis of sparse recovery based on constrained minimal singular values". In: *IEEE Transactions on Signal Processing* 59.12 (2011), pp. 5734–5745.

[TN12] Tang, G. and Nehorai, A. "The stability of low-rank matrix reconstruction: a constrained singular value view". In: *IEEE Transactions on Information Theory* 58.9 (2012), pp. 6079–6092.

[Tsy09] Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer, 2009.

[TT11] Tibshirani, R. J. and Taylor, J. "The Solution Path of the Generalized Lasso". In: *Annals of statistics* 39.3 (2011), pp. 1335–1371.

[Tuk58] Tukey, J. W. "Bias and confidence in not-quite large samples". In: *Annals of Mathematical Statistics*. Vol. 29. 2. 1958, pp. 614–614.

[Ush99] Ushakov, N. G. *Selected topics in characteristic functions*. Walter de Gruyter, 1999.

[Van+14] Van de Geer, S. et al. "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *The Annals of Statistics* 42.3 (2014), pp. 1166–1202.

[Ver12] Vershynin, R. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed Sensing: Theory and Applications*. Ed. by Eldar, Y. C. and Kutyniok, G. 2012.

[VF08] Van Den Berg, E. and Friedlander, M. P. “Probing the Pareto frontier for basis pursuit solutions”. In: *SIAM Journal on Scientific Computing* 31.2 (2008), pp. 890–912.

[VW96] Vaart, A. van der and Wellner, J. *Weak convergence and empirical processes.* Springer Verlag, 1996.

[Wan+14] Wan, J. et al. “Pairwise Costs in Semisupervised Discriminant Analysis for Face Recognition”. In: *IEEE Transactions on Information Forensics and Security* 9.10 (2014), pp. 1569–1580.

[War09] Ward, R. “Compressed sensing with cross validation”. In: *IEEE Transactions on Information Theory* 55.12 (2009), pp. 5773–5782.

[Was06] Wasserman, L. *All of nonparametric statistics.* Springer, 2006.

[Wig58] Wigner, E. P. “On the distribution of the roots of certain symmetric matrices”. In: *Annals of Mathematics* (1958), pp. 325–327.

[ZH05] Zou, H. and Hastie, T. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[Zol86] Zolotarev, V. *One-Dimensional Stable Distributions.* Vol. 65. American Mathematical Society, 1986.

[ZZ14] Zhang, C.-H. and Zhang, S. S. “Confidence intervals for low dimensional parameters in high dimensional linear models”. In: *Journal of the Royal Statistical Society: Series B* 76.1 (2014), pp. 217–242.

Appendix A

Proofs for Chapter 2

A.1 Proof of Theorem 2.1

Proof. Due to line (2.4) and Lemma 8.8 in B&F 1981,

$$d_2^2(\Psi_\rho(F_0;c), \Phi_\rho(\widehat{F};c)) = d_2^2\big(\Phi_\rho(F_0;c), \Phi_\rho(\widehat{F};c)\big) + (c^\top \delta(X))^2. \tag{A.1}$$

If $\varepsilon^* \in \mathbb{R}^n$ is a random vector whose entries are drawn i.i.d. from $\widehat{F}$, then the definition of Φ_ρ gives the matching relations

$$\begin{aligned} \Phi_\rho(F_0;c) &= \mathcal{L}(c^\top(X^\top X + \rho)^{-1} X^\top \varepsilon \,\big|\, X) \\ \Phi_\rho(\widehat{F};c) &= \mathcal{L}(c^\top(X^\top X + \rho)^{-1} X^\top \varepsilon^* \,\big|\, X). \end{aligned} \tag{A.2}$$

To make use of these relations, we apply Lemma 8.9 in B&F 1981, which implies that if $w \in \mathbb{R}^n$ is a generic deterministic vector, and if $U = (U_1, \dots, U_n)$ and $V = (V_1, \dots, V_n)$ are random vectors with i.i.d. entries, then

$$d_2^2(w^\top U, w^\top V) \leq \|w\|_2^2 \cdot d_2(U_1, V_1)^2.$$

Therefore,

$$\begin{aligned} d_2^2\big(\Phi_\rho(F_0), \Phi_\rho(\widehat{F})\big) &\leq \|c^\top (X^\top X + \rho)^{-1} X^\top\|_2^2 \cdot d_2^2(\varepsilon_1, \varepsilon_1^*) \\ &= \tfrac{1}{\sigma^2} \cdot v_\rho(X;c) \cdot d_2^2(F_0, \widehat{F}). \end{aligned} \tag{A.3}$$

Combining this with line (A.1) and dividing through by v_ρ proves the claim. □

A.2 Proof of Theorem 2.2

Proof. By the triangle inequality,

$$d_2^2(\widehat{F}, F_0) \leq 2\, d_2^2(\widehat{F}, F_n) + 2\, d_2^2(F_n, F_0). \tag{A.4}$$

Let $\tilde{F}_n$ be the (uncentered) empirical distribution of the residuals $\widehat{e}$ of $\widehat{\beta}$, which places mass $1/n$ at each value $\widehat{e}_i$, for $i = 1, \dots, n$. The proofs of Lemmas 2.1 and 2.2 in Freedman 1981, show that

$$\begin{aligned} \mathbb{E}\left[d_2(\widehat{F}, F_n)^2 \,\middle|\, X\right] &\leq \mathbb{E}\big[\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \varepsilon_i\big)^2\big] + \mathbb{E}\big[\, d_2(\tilde{F}_n, F_n)^2 \,\big|\, X\big] \\ &\leq \tfrac{1}{n}\sigma^2 + \tfrac{1}{n}\mathbb{E}\big[\|\widehat{e} - \varepsilon\|_2^2 \,\big|\, X\big] \\ &= \tfrac{1}{n}\sigma^2 + \tfrac{1}{n}\mathbb{E}\big[\|X(\beta - \widehat{\beta})\|_2^2 \,\big|\, X\big], \end{aligned} \tag{A.5}$$

where we have used the algebraic identity $\widehat{e} - \varepsilon = X(\beta - \widehat{\beta})$, which holds for any estimator $\widehat{\beta}$. This completes the proof. □

A.3 Proof of Theorem 2.3

Proof. We begin with a simple bias-variance decomposition,

$$\text{mspe}(\widehat{\beta}_\varrho|X) = \tfrac{1}{n}\mathbb{E}\Big[\big\|X\big(\widehat{\beta}_\varrho - \mathbb{E}\big[\widehat{\beta}_\varrho|X\big]\big)\big\|_2^2 \,\Big|\, X\Big] + \tfrac{1}{n}\big\|X(\mathbb{E}\big[\widehat{\beta}_\varrho|X\big] - \beta)\big\|_2^2. \tag{A.6}$$

We will handle the bias and variance terms separately. To consider the bias term, note that $\mathbb{E}[\widehat{\beta}_\rho|X] - \beta = B\beta$, where

$$B = (X^\top X + \varrho I_{p\times p})^{-1}X^\top X - I_{p\times p}.$$

Hence,

$$\begin{aligned} \tfrac{1}{n}\|X(\mathbb{E}[\widehat{\beta}_\varrho|X] - \beta)\|_2^2 &= \tfrac{1}{n}\|XB\beta\|_2^2 \\ &= \beta^\top B(\tfrac{1}{n}X^\top X)B\beta. \end{aligned} \tag{A.7}$$

If we let $l_i = \lambda_i(\frac{1}{n}X^\top X)$, then the eigenvalues of $B(\frac{1}{n}X^\top X)B$ are of the form $\mu_i := \frac{l_i(\varrho/n)^2}{(l_i+\varrho/n)^2}$. In particular, it is simple to check[1] that $\max_i \mu_i \asymp \varrho/n$ whenever $\theta \leq \nu$, and so

$$\beta^\top B(\tfrac{1}{n}X^\top X)B\beta \lesssim \tfrac{\varrho}{n}\|\beta\|_2^2 = n^{-\theta}\|\beta\|_2^2. \tag{A.8}$$

Note that this bound is tight, since it is achieved whenever β is parallel to the top eigenvector of $B(\frac{1}{n}X^\top X)B$.

To consider the variance term, note that $\widehat{\beta}_\varrho - \mathbb{E}[\widehat{\beta}_\varrho|X] = (X^\top X + \varrho I_{p\times p})^{-1}X^\top\varepsilon$, and so

$$\begin{aligned} \tfrac{1}{n}\mathbb{E}\Big[\big\|X(\widehat{\beta}_\varrho - \mathbb{E}[\widehat{\beta}_\varrho|X])\big\|_2^2 \,\Big|\, X\Big] &= \tfrac{1}{n}\,\text{tr}\left(\left(X^\top X\big(X^\top X + \varrho I_{p\times p}\big)^{-1}\right)^2\right) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n\wedge p}\big(\tfrac{l_i}{l_i+\varrho/n}\big)^2. \end{aligned} \tag{A.9}$$

[1] Note that if $t \in \mathbb{R}$ and $f(t) := \frac{t(\varrho/n)^2}{(t+\varrho/n)^2}$, then f is maximized at $t = \varrho/n$. Also, if $\theta \leq \nu$, then there at least one l_i that scales at the rate of ϱ/n.

It is natural to decompose the sum in terms of the index set

$$\mathcal{I}(n) := \{i \in \{1, \dots, n \wedge p\} : l_i \geq \varrho/n\}, \tag{A.10}$$

which satisfies $|\mathcal{I}(n)| \asymp n^{\frac{\theta}{\nu}}$. We will bound the variance term in two complementary cases; either $\nu > 1/2$ or $\nu \leq 1/2$. First assume $\nu > 1/2$. Then,

$$\frac{1}{n}\sum_{i=1}^{p}\big(\tfrac{l_i}{l_i+\varrho/n}\big)^2 = \frac{1}{n}\sum_{i\in\mathcal{I}(n)}\big(\tfrac{l_i}{l_i+\varrho/n}\big)^2 + \frac{1}{n}\sum_{i\notin\mathcal{I}(n)}\big(\tfrac{l_i}{l_i+\varrho/n}\big)^2 \tag{A.11}$$

$$\lesssim \frac{1}{n}\,|\mathcal{I}(n)| + \frac{1}{n}\int_{|\mathcal{I}(n)|}^{n\wedge p}\frac{x^{-2\nu}}{(\varrho/n)^2}dx \tag{A.12}$$

$$\lesssim n^{-1}\Big(n^{\frac{\theta}{\nu}} + n^{2\theta}\cdot(|\mathcal{I}(n)|)^{1-2\nu}\Big) \quad \text{using } \nu > \tfrac{1}{2} \tag{A.13}$$

$$\asymp n^{-1}\Big(n^{\frac{\theta}{\nu}} + n^{2\theta}\cdot(n^{\frac{\theta}{\nu}})^{(1-2\nu)}\Big) \tag{A.14}$$

$$= 2n^{\frac{\theta-\nu}{\nu}}. \tag{A.15}$$

To see that this upper bound is tight, note that in line (A.11), we can use the term-wise lower bounds

$$\big(\tfrac{l_i}{l_i+\varrho/n}\big)^2 \geq \begin{cases} \frac{1}{4} & \text{if } i \in \mathcal{I}(n) \\ \frac{1}{4}\frac{l_i^2}{(\varrho/n)^2} & \text{if } i \notin \mathcal{I}(n), \end{cases} \tag{A.16}$$

and then apply an integral approximation from below (which leads to the same rate). Combining the bias and variance pieces, we have shown that

$$\tfrac{1}{n}\mathbb{E}\|X(\widehat{\beta}_\varrho - \beta)\|_2^2 \lesssim n^{\frac{\theta-\nu}{\nu}} + n^{-\theta} \qquad \text{if} \quad \nu > \tfrac{1}{2}.$$

The bound is optimized when the two terms on the right side have the same rate, which leads to the choice $\theta = \frac{\nu}{\nu+1}$.

In the case where $\nu \in (0, \frac{1}{2})$, the calculation proceeds in the same way up to line (A.13), where we obtain the bound

$$\frac{1}{n}\sum_{i=1}^{n\wedge p}\big(\tfrac{l_i}{l_i+\varrho/n}\big)^2 \lesssim n^{-1}\Big(n^{\frac{\theta}{\nu}} + n^{2\theta}\cdot n^{1-2\nu}\Big) \tag{A.17}$$

$$= n^{\frac{\theta-\nu}{\nu}} + n^{2(\theta-\nu)}. \tag{A.18}$$

This bound is also tight due to the same reasoning as above. Note that in order for the bound to tend to 0 as $n \to \infty$, we must choose $\theta < \nu$. Furthermore, since we are working under the assumption $\nu \in (0, \frac{1}{2})$, it follows that the right side of line (A.18) has rate equal to $n^{2(\theta-\nu)}$. Combining the rates for the bias and variance shows that

$$\tfrac{1}{n}\mathbb{E}\|X(\widehat{\beta}_\varrho - \beta)\|_2^2 \lesssim n^{2(\theta-\nu)} + n^{-\theta} \qquad \text{if} \quad \nu \in (0, \tfrac{1}{2}).$$

The bound is optimized when the two terms on the right side have the same rate, which leads to the choice $\theta = \frac{2\nu}{3}$. □

A.4 Proof of Lemma 2.2

The proof is split up into three pieces, corresponding to parts (i), (ii), and (iii) in the statement of the result.

The bias inequality (2.14)

We prove inequality (2.14) by combining Lemmas A.1 and A.3 below.

Lemma A.1. *Assume the conditions of Lemma 2.2. For each $i \in \{1, \dots, n\}$, there are independent random vectors $u_i(X), w(X) \in \mathbb{R}^p$ such that the random variable $X_i^\top \delta(X)$ can be represented algebraically as*

$$b_\rho(X; X_i) = X_i^\top \delta(X) = u_i(X)^\top w(X).$$

Here, the vectors $u_i(X)$ can be represented in law as

$$u_i(X) \overset{\mathcal{L}}{=} \tfrac{1}{\|z\|_2} \Pi_p(z), \tag{A.19}$$

where $z \in \mathbb{R}^n$ is a standard Gaussian vector, and $\Pi_p(z) := (z_1, \dots, z_p)$. Also, the vector $w(X)$ satisfies the bound $\|w(X)\|_2^2 \leq \frac{\rho}{4}\|\beta\|_2^2$ almost surely.

Proof. To fix notation, we write $X^\top = \Sigma^{1/2} Z^\top$ where $Z^\top \in \mathbb{R}^{p \times n}$ is a standard Gaussian matrix. Recall that $\delta(X) = B\beta$, where

$$B = I_{p\times p} - (X^\top X + \rho I_{p\times p})^{-1} X^\top X.$$

Let $Z = HLG^\top$ be a signed s.v.d. for Z, as defined in Section A.5 in Appendix A, where $H \in \mathbb{R}^{n\times p}$, $L \in \mathbb{R}^{p\times p}$, and $G \in \mathbb{R}^{p\times p}$. Now define $u_i(X)$ and $w(X)$ according to

$$X_i^\top \delta(X) = e_i^\top X B\beta = e_i^\top Z \Sigma^{1/2} B\beta = \underbrace{e_i^\top H}_{=:\, u_i(X)^\top} \; \underbrace{LG^\top \Sigma^{1/2} B\beta}_{=:\, w(X)}. \tag{A.20}$$

From Lemma A.15 in that appendix, the rows $e_i^\top H$ can be represented in distribution as $\frac{1}{\|z\|_2}\Pi_p(z)$. The same lemma also shows that the three matrices H, L, and G are independent.

Hence, to show that $u_i(X)$ and $w(X)$ are independent, it suffices to show that $w(X)$ is a function only of G and L. In turn, it is enough to show that B is a function only of G and L. But this is simple, because B is a function only of the matrix $X^\top X$, which may be written as

$$X^\top X = \Sigma^{1/2} Z^\top Z \Sigma^{1/2} = \Sigma^{1/2} G L^2 G^\top \Sigma^{1/2}. \tag{A.21}$$

It remains to show that $\|w(X)\|_2^2 \leq \frac{\rho}{4}\|\beta\|_2^2$ almost surely. Combining the definition of $w(X)$ with line (A.21) gives

$$\|w(X)\|_2^2 = \beta^\top \big(BX^\top XB\big)\beta. \tag{A.22}$$

The eigenvalues of $BX^\top XB$ are of the form $\mu_i := n\frac{(\rho/n)^2 l_i}{(l_i+\rho/n)^2}$ where $l_i = \lambda_i(\frac{1}{n}X^\top X)$, and it is simple to check that the inequality $\max_i \mu_i \le \frac{\rho}{4}$ holds for every realization of X. □

Before proceeding to the second portion of the proof of inequality (2.14), we record some well-known tail bounds for Gaussian quadratic forms due to Laurent and Massart [LM00], which will be useful at various points later on.

Lemma A.2 (Laurent & Massart, 2001)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a fixed symmetric matrix, and let $z \in \mathbb{R}^n$ be a standard Gaussian vector. Then, for every $t > 0$,*

$$\mathbb{P}\Big[z^\top Az \ge \operatorname{tr}(A) + 2\,|||A|||_F\,\sqrt{t} + 2\,|||A|||_{\text{op}}\,t\Big] \le \exp(-t) \tag{A.23}$$

and

$$\mathbb{P}\Big[z^\top Az \le \operatorname{tr}(A) - 2\,|||A|||_F\,\sqrt{t}\Big] \le \exp(-t). \tag{A.24}$$

The next lemma completes the proof of inequality (2.14).

Lemma A.3. *Assume the conditions of Lemma 2.2, and let $\tau > 0$ be a constant. Then for every $n \ge 1$, the following event holds with probability at least $1 - n^{-\tau} - ne^{-n/16}$,*

$$\max_{1\le i\le n} b_\rho^2(X;X_i) \;\le 5\|\beta\|_2^2\cdot n^{-\gamma}\cdot(\tau+1)\log(n+2). \tag{A.25}$$

Proof. Applying the representation for $b_\rho(X;X_i)$ given in Lemma A.1, there is a standard Gaussian vector $z \in \mathbb{R}^n$, such that $u_i(X) \overset{\mathcal{L}}{=} \Pi_p(z)/\|z\|_2$. Consequently,

$$b_\rho^2(X;X_i) \overset{\mathcal{L}}{=} \tfrac{1}{\|z\|_2^2}\cdot\Pi_p(z)\Big(w(X)w(X)^\top\Big)\Pi_p(z), \tag{A.26}$$

where we may take Z and $w(X)$ to be independent by the same lemma. Using Lemma A.2 on Gaussian quadratic forms, as well as the fact that $\|w(X)\|_2^2 \le \frac{\rho}{4}\|\beta\|_2^2$ almost surely, we have for all $t > 0$,

$$\mathbb{P}\left[\Pi_p(z)^\top\Big(w(X)w(X)^\top\Big)\Pi_p(z) \ge \tfrac{\rho}{4}\|\beta\|_2^2\big(1+2\sqrt{t}+2t\big)\;\middle|\; w(X)\right] \le \exp(-t). \tag{A.27}$$

The same lemma also implies that for all $t' \in (0, \frac{1}{4})$,

$$\mathbb{P}\Big[\tfrac{1}{\|z\|_2^2} \ge \tfrac{1}{(1-2\sqrt{t'})n}\Big] \le \exp(-nt'). \tag{A.28}$$

Now, we combine the bounds by integrating out $w(X)$ in line (A.27) and choosing $t' = 1/16$ in line (A.28). Taking a union bound, we conclude that for any $t > 0$, and any fixed $i = 1, \dots, n$,

$$\mathbb{P}\left[b_\rho^2(X;X_i) \;\le\; \tfrac{\rho}{n}\cdot\|\beta\|_2^2\cdot\tfrac{1}{2}\big(1+2\sqrt{t}+2t\big)\right] \ge 1 - e^{-t} - e^{-n/16}. \tag{A.29}$$

Finally, another union bound shows that the maximum of the $b_\rho(X, X_i^\top)$ satisfies

$$\mathbb{P}\Bigg[\max_{1\leq i\leq n} b_\rho^2(X;X_i) \;\leq\; \tfrac{\rho}{n}\cdot\|\beta\|_2^2\cdot\tfrac{1}{2}(1+2\sqrt{t}+2t)\Bigg] \geq 1-e^{-t+\log(n)}-ne^{-n/16}, \tag{A.30}$$

which implies the stated result after choosing $t=(\tau+1)\log(n+2)$, and noting that since $t\geq 1$, we have $\frac{1}{2}(1+2\sqrt{t}+2t)\leq 5t=5(\tau+1)\log(n+2)$, as well as $e^{-t+\log(n)}\leq e^{-\tau\log(n+2)}\leq n^{-\tau}$ for every $n\geq 1$. □

The variance inequality (2.15)

The following "representation lemma" will serve as the basis for controlling the variance $v_\rho(X;X_i)=\sigma^2\|X_i^\top(X^\top X+\rho I_{p\times p})^{-1}X^\top\|_2^2$.

Lemma A.4. *Assume the conditions of Lemma 2.2. For each $i\in\{1,\dots,n\}$, there is a random vector $v_i(X)\in\mathbb{R}^p$ and a random matrix $M(X)\in\mathbb{R}^{p\times p}$ that are independent and satisfy the algebraic relation*

$$\|X_i^\top(X^\top X+\rho I_{p\times p})^{-1}X^\top\|_2^2=v_i(X)^\top M(X)v_i(X).$$

Here, the vector $v_i(X)$ can be represented in law as

$$v_i(X)\overset{\mathcal{L}}{=}\tfrac{1}{\|z\|_2}\Pi_p(z), \tag{A.31}$$

where $z\in\mathbb{R}^n$ is a standard Gaussian vector and $\Pi_p(z)=(z_1,\dots,z_p)$. Also, the matrix $M(X)$ satisfies the algebraic relation

$$\operatorname{tr}(M(X))=\|X(X^\top X+\rho I_{p\times p})^{-1}X^\top\|_F^2. \tag{A.32}$$

An explicit formula for $M(X)$ is given below.

Proof. Define the matrix $A:=(X^\top X+\rho I_{p\times p})^{-1}X^\top X(X^\top X+\rho I_{p\times p})^{-1}$. Then,

$$\|X_i^\top(X^\top X+\rho I_{p\times p})^{-1}X^\top\|_2^2=e_i^\top XAX^\top e_i. \tag{A.33}$$

Using the notation in the proof of the previous lemma, let $X=Z\Sigma^{1/2}$ where $Z\in\mathbb{R}^{n\times p}$ is a standard Gaussian random matrix. Furthermore, let $Z=HLG^\top$ be a signed s.v.d. for Z, as defined in Section A.5. Then, we define $v_i(X)$ and $M(X)$ according to

$$e_i^\top XAX^\top e_i=\underbrace{e_i^\top H}_{=:v_i(X)^\top}\;\underbrace{LG^\top\Sigma^{1/2}A\Sigma^{1/2}GL^\top}_{=:M(X)}H^\top e_i. \tag{A.34}$$

Some algebra shows that $M(X)$ satisfies the relation (A.32). As in the proof of Lemma A.1, the argument is completed using two properties of the signed s.v.d. of a standard Gaussian matrix: The rows of H can be represented as $\Pi_p(z)/\|z\|_2$ where $z \in \mathbb{R}^n$ is a standard Gaussian vector, and the matrices H, L, and $G^\top$ are independent. (See Lemma A.15 in Section A.5.) To show that $v_i(X)$ and $M(X)$ are independent, first note that $v_i(X)$ only depends on H. Also, it is simple to check that $M(X)$ only depends on G and L, because A is a function only of $X^\top X = \Sigma^{1/2} G L^2 G^\top \Sigma^{1/2}$.)

□

Concentration of the variance and bounds on its expected value

Due to Lemma A.4, for each $i = 1, \dots, n$, we have the representation

$$v_\rho(X; X_i) \overset{\mathcal{L}}{=} \tfrac{1}{\|z\|_2^2} \Pi_p(z)^\top M(X) \Pi_p(z), \tag{A.35}$$

where $z \in \mathbb{R}^n$ is a standard Gaussian vector, independent of $M(X)$. Conditionally on $M(X)$, the quadratic form $\Pi_p(z)^\top M(X) \Pi_p(z)$ concentrates around $\mathrm{tr}(M(X))$ by Lemma A.2. The same lemma also implies that $\|z\|_2^2$ concentrates around n. In the next three subsections, we will show that $\sqrt{\mathrm{tr}(M(X))}$ concentrates around its expected value, and obtain upper and lower bounds on the expected value. We will need two-sided bounds in preparation for Theorem 2.4.

Concentration of $\sqrt{\mathrm{tr}(M(X))}$

Lemma A.5. *Assume the conditions of Lemma 2.2. Then for every $t > 0$, and every $n \geq 1$,*

$$\mathbb{P}\left[\left| \sqrt{\mathrm{tr}(M(X))} - \mathbb{E}\sqrt{\mathrm{tr}(M(X))} \right| \geq t \right] \leq 2 \exp(-\tfrac{64}{54} \tfrac{n^{1-\gamma} t^2}{\|\Sigma\|_{\mathrm{op}}}). \tag{A.36}$$

Proof. We will show that $\sqrt{\mathrm{tr}(M(X))}$ is a Lipschitz function of a standard Gaussian matrix. Define the function $g_\rho : \mathbb{R}_+ \to [0,1]$ by $g_\rho(s) = \frac{s^2}{s^2 + (\rho/n)}$, which satisfies the Lipschitz condition

$$|g_\rho(s) - g_\rho(s')| \leq \mathfrak{L}_n |s - s'|,$$

for all $s, s' \geq 0$, where $\mathfrak{L}_n := \frac{3\sqrt{3}}{8} \frac{1}{\sqrt{\rho/n}}$.

If $\sigma(A) = (\sigma_1(A), \dots, \sigma_k(A))$ denotes the vector of singular values of a rank k matrix A, then we define g_ρ to act on $\sigma(A)$ component-wise, i.e. $g_\rho(\sigma(A)) = (g_\rho(\sigma_1(A)), \dots, g_\rho(\sigma_k(A)))$. Recall from Lemma A.4 that

$$\sqrt{\mathrm{tr}(M(X))} = \|X(X^\top X + \rho I_{p\times p})^{-1} X^\top\|_F \tag{A.37}$$

and note that the ith singular value of the matrix $X(X^\top X + \rho I_{p\times p})^{-1} X^\top$ is given by $g_\rho(\sigma_i(\frac{1}{\sqrt{n}} X))$. Viewing the Frobenius norm of a matrix as the ℓ_2 norm of its singular values, we have

$$\sqrt{\mathrm{tr}(M(X))} = \|g_\rho(\sigma(\tfrac{1}{\sqrt{n}} X))\|_2. \tag{A.38}$$

Write $X^\top = \Sigma^{1/2} Z^\top$ for a standard Gaussian matrix $Z \in \mathbb{R}^{n\times p}$, and let $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ be defined according to

$$f(Z) := \sqrt{\operatorname{tr}(M(X))}.$$

We claim that f is Lipschitz with respect to the Frobenius norm. Let $W^\top \in \mathbb{R}^{p\times n}$ be a generic matrix, and put $A = \frac{1}{\sqrt{n}}\Sigma^{1/2}Z^\top$ and $B = \frac{1}{\sqrt{n}}\Sigma^{1/2}W^\top$. Then,

$$\begin{aligned}
|f(Z) - f(W)| &= \Big| \|g_\rho(\sigma(A))\|_2 - \|g_\rho(\sigma(B))\|_2 \Big| && \text{(A.39)}\\
&\le \big\| g_\rho(\sigma(A)) - g_\rho(\sigma(B)) \big\|_2 && \text{(A.40)}\\
&\le \mathfrak{L}_n \|\sigma(A) - \sigma(B)\|_2 && \text{(A.41)}\\
&\le \mathfrak{L}_n \|A - B\|_F \qquad \text{(Weilandt-Hoffman)} && \text{(A.42)}\\
&= \mathfrak{L}_n \big\| \tfrac{1}{\sqrt{n}} \Sigma^{1/2} \big(Z^\top - W^\top\big) \big\|_F && \text{(A.43)}\\
&\le \tfrac{\mathfrak{L}_n}{\sqrt{n}} \sqrt{\|\Sigma\|_{\text{op}}} \cdot \big\| Z^\top - W^\top \big\|_F, && \text{(A.44)}
\end{aligned}$$

where we have used a version of the Weilandt-Hoffman inequality for singular values [HJ91, p.186], as well as the inequality $\|M_1M_2\|_F \le \|M_1\|_{\text{op}}\|M_2\|_F$, which holds for any square matrix M_1 that is compatible with M_2. (See Lemma A.10 in Appendix A.5.) The statement of the lemma now follows from the Gaussian concentration inequality. (See Lemma A.13 in Appendix A.5). □

Upper bound on $\mathbb{E}\sqrt{\operatorname{tr}(M(X))}$

Lemma A.6. *Assume the conditions of Lemma 2.2. Then, the matrix $M(X)$ satisfies*

$$\mathbb{E}\sqrt{\operatorname{tr}(M(X))} \lesssim \begin{cases} n^{(\gamma-\eta)+\frac{1}{2}} & \text{if } \eta \in (0, \frac{1}{2}) \\ n^{\frac{\gamma}{2\eta}} & \text{if } \eta > \frac{1}{2}. \end{cases} \tag{A.45}$$

Proof. By Jensen's inequality, it is enough to bound $\sqrt{\mathbb{E}[\operatorname{tr}(M(X))]}$ from above. Define the univariate function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ by $\psi(s) := \frac{s}{(\sqrt{s}+\rho/n)^2}$, and observe that

$$\begin{aligned}
\operatorname{tr}(M(X)) &= \operatorname{tr}\Big(\Big((X^\top X + \rho I_{p\times p})^{-1} X^\top X \Big)^2 \Big)\\
&= \sum_{i=1}^{p} \tfrac{\lambda_i^2(\widehat{\Sigma})}{(\lambda_i(\widehat{\Sigma})+\rho/n)^2}\\
&= \sum_{i=1}^{p} \psi(\lambda_i(\widehat{\Sigma}^2))\\
&= \operatorname{tr}\big(\psi(\widehat{\Sigma}^2)\big).
\end{aligned} \tag{A.46}$$

Here where we use the "operator calculus" notation $\psi(A) = U\psi(D)U^\top$ where A is a symmetric matrix with spectral decomposition $A = UDU^\top$, and $\psi(D)$ is the diagonal matrix whose ith diagonal entry is $\psi(D_{ii})$. It is simple to check that ψ is a concave, and so $\text{tr}(\psi(\widehat{\Sigma}^2))$ is a concave matrix functional of $\widehat{\Sigma}^2$ by Lemma A.11 in Section A.5 of Appendix A. Therefore, Jensen's inequality implies

$$\begin{aligned}\mathbb{E}[\text{tr}(M(X))] &\leq \text{tr}(\psi(\mathbb{E}[\widehat{\Sigma}^2])) \\ &= \textstyle\sum_{i=1}^p \psi(\lambda_i(\mathfrak{S})),\end{aligned} \tag{A.47}$$

where we define the matrix $\mathfrak{S} := \mathbb{E}[\widehat{\Sigma}^2]$. Since X is Gaussian, $\widehat{\Sigma}$ is a Wishart matrix up to scaling, and so Lemma A.14 in Appendix A.5 shows that this expectation may be evaluated exactly as

$$\mathfrak{S} = (1 + \tfrac{1}{n})\Sigma^2 + \tfrac{\text{tr}(\Sigma)}{n}\Sigma. \tag{A.48}$$

We will now use this relation to apply an integral approximation to the right side of line (A.47). Clearly, the eigenvalues of $\mathfrak{S}$ are given by

$$\begin{aligned}\lambda_i(\mathfrak{S}) &= (1 + \tfrac{1}{n})\lambda_i^2(\Sigma) + \tfrac{\text{tr}(\Sigma)}{n}\lambda_i(\Sigma) \\ &\asymp i^{-2\eta} + \tfrac{\text{tr}(\Sigma)}{n} i^{-\eta}.\end{aligned} \tag{A.49}$$

Let $r \in (0,1)$ be a constant to be specified later. On the set of indices $1 \leq i \leq \lceil n^r \rceil$ we use the bound $\psi(\lambda_i(\mathfrak{S})) \leq 1$, and on the set of indices $i > \lceil n^r \rceil$ we use the bound $\psi(\lambda_i(\mathfrak{S})) \leq \frac{1}{(\rho/n)^2}\lambda_i(\mathfrak{S})$. Recalling the assumption $\rho/n = n^{-\gamma}$, we may decompose the inequality (A.47) as[2]

$$\mathbb{E}[\text{tr}(M(X))] \leq \sum_{i=1}^{\lceil n^r \rceil} \psi(\lambda_i(\mathfrak{S})) + \sum_{i=\lceil n^r \rceil + 1}^{p} \psi(\lambda_i(\mathfrak{S})) \tag{A.50}$$

$$\lesssim n^r + n^{2\gamma} \int_{n^r}^{p} \big(x^{-2\eta} + \tfrac{\text{tr}(\Sigma)}{n} x^{-\eta}\big) dx \tag{A.51}$$

$$=: n^r + n^{2\gamma} h_n(\eta, r). \tag{A.52}$$

where the function h_n is defined in the last line. The bound is optimized when the two terms on the right are of the same order; i.e. when r solves the rate equation

$$n^r \asymp n^{2\gamma} h_n(\eta, r). \tag{A.53}$$

Noting that

$$\tfrac{\text{tr}(\Sigma)}{n} \asymp \begin{cases} n^{-\eta} & \text{if } \eta \in (0,1) \\ n^{-1} & \text{if } \eta > 1, \end{cases} \tag{A.54}$$

[2] Note that if $p/n \asymp 1$, it is still possible that $n^r > p$ for small values of n. Since we want $n^r \leq p$ for the integral in line (A.51), Lemma 2.2 is stated for "all large n".

the quantity $h_n(\eta, r)$ may be computed directly as

$$h_n(\eta, r) \asymp \begin{cases} n^{1-2\eta} & \text{if } \ \eta \in (0, \frac{1}{2}), \\ n^{r(1-2\eta)} & \text{if } \ \eta \in (\frac{1}{2}, 1), \\ n^{r(1-2\eta)} + n^{r(1-\eta)-1} & \text{if } \ \eta > 1. \end{cases} \tag{A.55}$$

If we let $r = r_*(\eta, \gamma)$ denote the solution of the rate equation (A.53), then some calculation shows that under the assumption $\gamma \in (0, 1)$,

$$r_*(\eta, \gamma) = \begin{cases} 2(\gamma - \eta) + 1 & \text{if } \ \eta \in (0, \frac{1}{2}), \\ \frac{\gamma}{\eta} & \text{if } \ \eta > \frac{1}{2}. \end{cases} \tag{A.56}$$

When $\eta \in (0, 1)$ this is straightforward. To show the details for $\eta > 1$, note that the rate equation (A.53) may be written as

$$n^r \asymp n^{2\gamma + r(1-2\eta)} + n^{2\gamma + r(1-\eta) - 1}, \tag{A.57}$$

which is the same as

$$1 \asymp n^{2(\gamma - \eta r)} + n^{2\gamma - \eta r - 1}. \tag{A.58}$$

In order for both terms on the right to be $\mathcal{O}(1)$, the number r must satisfy the constraints

$$r \geq \tfrac{\gamma}{\eta}, \tag{A.59}$$

$$r \geq \tfrac{\gamma}{\eta} + \tfrac{\gamma - 1}{\eta}. \tag{A.60}$$

Since Lemma 2.2 assumes $\gamma \in (0, 1)$, only the first constraint matters. Furthermore, when $r \geq \frac{\gamma}{\eta}$, the second term in line (A.58) is $o(1)$, and we are reduced to choosing r so that $1 \asymp n^{2(\gamma - \eta r)}$, which gives $r = r_*(\eta, \gamma) = \frac{\gamma}{\eta}$. Substituting this value into line (A.52) completes the proof. (Note from the discussion preceding line (A.52) that r must lie in the interval $(0, 1)$, and this requires $\gamma/\eta < 1$, which explains the assumption $\gamma < \min\{\eta, 1\}$ in Lemma 2.2.)

Lower bound on $\mathbb{E}\sqrt{\mathrm{tr}(M(X))}$

Lemma A.7. *Assume the conditions of Lemma 2.2. Then, the matrix $M(X)$ satisfies*

$$\mathbb{E}\sqrt{\mathrm{tr}(M(X))} \gtrsim n^{\frac{\gamma}{2\eta}}. \tag{A.61}$$

Proof. The variable $\sqrt{\mathrm{tr}(M(X))}$ may be written as $\|X^\top X(X^\top X + \rho I_{p\times p})^{-1}\|_F$. Since the Frobenius norm is a convex matrix functional, Jensen's inequality implies

$$\begin{aligned} \mathbb{E}\sqrt{\mathrm{tr}(M(X))} &\geq \Big\|\mathbb{E}\Big[X^\top X(X^\top X + \rho I_{p\times p})^{-1}\Big]\Big\|_F \\ &= \Big\|\mathbb{E}\Big[\big(I_{p\times p} + \tfrac{\rho}{n}\widehat{\Sigma}^{-1}\big)^{-1}\Big]\Big\|_F, \end{aligned} \tag{A.62}$$

where the last step follows algebraically with $\widehat{\Sigma} := \frac{1}{n}X^\top X$. If we define the univariate function $f : \mathbb{R}_+ \to \mathbb{R}_+$ by $f(s) = (1 + \frac{\rho}{n}s)^{-1}$, then last inequality is the same as

$$\mathbb{E}\sqrt{\operatorname{tr}(M(X))} \geq \big\|\mathbb{E}\big[f\big(\widehat{\Sigma}^{-1}\big)\big]\big\|_F. \tag{A.63}$$

It is a basic fact that f is operator convex on the domain of positive semidefinite matrices [Bha97, p.117]. This yields an operator version of Jensen's inequality with respect to the Loewner ordering (Lemma A.12 in Appendix A.5):

$$\mathbb{E}\big[f\big(\widehat{\Sigma}^{-1}\big)\big] \succeq f\big(\mathbb{E}\big[\widehat{\Sigma}^{-1}\big]\big). \tag{A.64}$$

Furthermore, if two matrices satisfy $A \succeq B \succeq 0$, then $\|A\|_F \geq \|B\|_F$ [HJ09, Corollary 7.7.4]. Using this fact, as well as the formula for the expected inverse of a Wishart matrix [Mui82, p. 97], we obtain

$$\begin{aligned}
\mathbb{E}\sqrt{\operatorname{tr}(M(X))} &\geq \big\|f\big(\mathbb{E}\big[\widehat{\Sigma}^{-1}\big]\big)\big\|_F \\
&= \big\|f\big(\tfrac{n}{n-p-1}\Sigma^{-1}\big)\big\|_F \\
&= \left(\sum_{i=1}^{p} \frac{1}{\big(1 + \frac{\rho}{n}\cdot\frac{n}{n-p-1}\lambda_i(\Sigma^{-1})\big)^2}\right)^{1/2} \\
&= \left(\sum_{i=1}^{p} \frac{\lambda_i^2(\Sigma)}{\big(\lambda_i(\Sigma) + \frac{\rho}{n}\cdot\frac{n}{n-p-1}\big)^2}\right)^{1/2}.
\end{aligned} \tag{A.65}$$

Define the index set $J = \big\{i \in \{1,\dots,p\} : \lambda_i(\Sigma) \geq \frac{\rho}{n}\frac{n}{n-p-1}\big\}$. For any $i \in J$, the ith summand in the previous line is at least $1/4$. Also, assumption **A2.6** that p/n is bounded strictly between 0 and 1, as well as the decay condition on the $\lambda_i(\Sigma)$, imply that $|J| \asymp n^{\gamma/\eta}$, which completes the proof. □

Putting the variance pieces together

Combining Lemmas A.5, A.6, and A.7 with the Gaussian concentration inequality (Lemma A.13 in Section A.5 of Appendix A) immediately gives the following result. (We choose t to be proportional to the relevant bound on $\mathbb{E}[\sqrt{\operatorname{tr}(M(X))}]$ in the Gaussian concentration inequality.)

Lemma A.8. *Assume the conditions of Lemma 2.2 and let* $\operatorname{tr}(M(X))$ *be as in line* (A.32). *Then, there are absolute constants* $\kappa_1, \kappa_2, \dots, \kappa_6 > 0$ *such that the following upper-tail bounds hold for all large* n*,*

$$\mathbb{P}\Big[\operatorname{tr}(M(X)) \geq \kappa_1 n^{2(\gamma-\eta)+1}\Big] \leq \exp(-\kappa_2 n^{2(1-\eta)+\gamma}), \qquad \textit{if}\ \ \eta \in (0, \tfrac{1}{2}), \tag{A.66}$$

and

$$\mathbb{P}\Big[\operatorname{tr}(M(X)) \geq \kappa_3 n^{\gamma/\eta}\Big] \leq \exp(-\kappa_4 n^{1+\frac{\gamma(1-\eta)}{\eta}}), \qquad \textit{if} \;\; \eta > \tfrac{1}{2}, \tag{A.67}$$

and the following lower-tail bound holds for all large n,

$$\mathbb{P}\Big[\operatorname{tr}(M(X)) \leq \kappa_5 n^{\gamma/\eta}\Big] \leq \exp(-\kappa_6 n^{1+\frac{\gamma(1-\eta)}{\eta}}), \qquad \textit{if} \;\; \eta > 0. \tag{A.68}$$

Remarks. Note that in order for the last two probabilities to be small for large values of $\eta > 0$, it is necessary that $\gamma < 1$, as assumed in Lemma 2.2. The next result completes the assembly of the results in this Subsection A.4. Although the first two bounds in Lemma A.9 are not necessary for the statement of Theorem 2.4, they show that the variance $v_\rho(X; X_i)$ tends 0 as $n \to \infty$ when $\gamma < \eta$, as assumed in Theorem 2.4. In other words, we imposed the assumption $\gamma < \eta$ so that confidence intervals based on $\Phi_\rho(\widehat{F}_\varrho; X_i)$ have width that tends to 0 asymptotically.

Lemma A.9. *Assume the conditions of Theorem 2.4 and let* $\operatorname{tr}(M(X))$ *be as in line* (A.32). *Assume $\gamma < \min\{\eta, 1\}$. Then, there are absolute constants $k_1, k_2, \dots, k_6 > 0$ such that the following upper-tail bounds hold for all large n,*

$$\mathbb{P}\Big[\max_{1\leq i\leq n} v_\rho(X; X_i) \leq k_1 n^{2(\gamma-\eta)}\Big] \geq 1 - 4n\exp(-k_2 n^{\frac{\gamma}{\eta}}), \qquad \textit{if} \;\; \eta \in (0, \tfrac{1}{2}) \tag{A.69}$$

and

$$\mathbb{P}\Big[\max_{1\leq i\leq n} v_\rho(X; X_i) \leq k_3 n^{\frac{\gamma}{\eta}-1}\Big] \geq 1 - 4n\exp(-k_4 n^{\frac{\gamma}{\eta}}), \qquad \textit{if} \;\; \eta > \tfrac{1}{2}, \tag{A.70}$$

and

$$\mathbb{P}\Big[\max_{1\leq i\leq n} \tfrac{1}{v_\rho(X;X_i)} \leq k_5 n^{1-\frac{\gamma}{\eta}}\Big] \geq 1 - 4n\exp(-k_6 n^{\frac{\gamma}{\eta}}), \qquad \textit{if} \;\; \eta > 0. \tag{A.71}$$

Proof. We only prove the last inequality (A.71), since the other two inequalities are proven in a similar way. By Lemma A.4, we have

$$v_\rho(X; X_i) \overset{\mathcal{L}}{=} \tfrac{1}{\|z\|_2^2} \Pi_p(z)^\top M(X) \Pi_p(z) \tag{A.72}$$

where $z \sim N(0, I_{p\times p})$ and $z \perp\!\!\!\perp M(X)$. To apply the lower-tail bound for Gaussian quadratic forms, note that Hölder's inequality implies $\|M(X)\|_F \leq \sqrt{\operatorname{tr}(M(X))}$ since $\|M(X)\|_{\text{op}} \leq 1$ almost surely. Therefore, letting $t = t' \operatorname{tr}(M(X))$ with $t' \in (0,1)$ in inequality (A.24) gives

$$\mathbb{P}\Big[\Pi_p(z)^\top M(X) \Pi_p(z) \geq (1 - 2\sqrt{t'}) \operatorname{tr}(M(X)) \,\Big|\, M(X)\Big] \geq 1 - \exp\big(-t' \operatorname{tr}(M(X))\big) \tag{A.73}$$

Next, observe that inequality (A.24) with $t = t' \cdot n$ for $t' \in (0,1)$ gives,

$$\mathbb{P}\Big[\|z\|_2^2 \leq (1 + 4\sqrt{t'})n\Big] \geq 1 - \exp(-t'n). \tag{A.74}$$

If we define the event

$$\mathcal{E}_1 := \Bigg\{ \frac{1}{\frac{1}{\|z\|_2^2} \Pi_p(z)^\top M(X) \Pi_p(z)} \leq \tfrac{1+4\sqrt{t'}}{(1-2\sqrt{t'})} \tfrac{n}{\operatorname{tr}(M(X))} \Bigg\} \tag{A.75}$$

then the previous two inequalities imply

$$\begin{aligned} \mathbb{P}\big[\mathcal{E}_1 \, \big| \, M(X)\big] &\geq 1 - \exp(-t' \operatorname{tr}(M(X))) - \exp(-t' \cdot n) \\ &\geq 1 - 2\exp(-t' \operatorname{tr}(M(X))), \end{aligned} \tag{A.76}$$

since $\operatorname{tr}(M(X)) \leq n$ almost surely. Next, let $\kappa_5, \kappa_6 > 0$ be as in the previous lemma, and define the event

$$\mathcal{E}_2 := \Big\{ \tfrac{n}{\operatorname{tr}(M(X))} \leq \tfrac{1}{\kappa_5} n^{1-\frac{\gamma}{\eta}} \Big\}, \tag{A.77}$$

which has probability $\mathbb{P}(\mathcal{E}_2) \geq 1 - \exp(-\kappa_6 n^{1+\frac{\gamma(1-\eta)}{\eta}})$.

We now put these items together. Starting with line (A.72), if we work on the intersection of $\mathcal{E}_1$ and $\mathcal{E}_2$, then for any fixed $i = 1, \dots, n$ we have

$$\begin{aligned} \mathbb{P}\Big[\tfrac{1}{v_\rho(X;X_i)} \leq \tfrac{1+4\sqrt{t'}}{(1-2\sqrt{t'})} \tfrac{1}{\kappa_5} n^{1-\frac{\gamma}{\eta}}\Big] &\geq \mathbb{E}\Big[1_{\mathcal{E}_1} \cdot 1_{\mathcal{E}_2}\Big] \\ &\geq 1 - \mathbb{E}\Big[1_{\mathcal{E}_1^c} + 1_{\mathcal{E}_2^c}\Big] \\ &= 1 - \mathbb{E}\big[\mathbb{E}\big[1_{\mathcal{E}_1^c} \, \big| \, M(X)\big]\big] - \mathbb{P}(\mathcal{E}_2^c) \\ &\geq 1 - 2\mathbb{E}\Big[\exp(-t' \operatorname{tr}(M(X))\Big] - \mathbb{P}(\mathcal{E}_2^c) \\ &= 1 - 2\mathbb{E}\Big[\exp(-t' \operatorname{tr}(M(X))) \cdot (1_{\mathcal{E}_2} + 1_{\mathcal{E}_2^c})\Big] - \mathbb{P}(\mathcal{E}_2^c) \\ &\geq 1 - \exp(-t' \kappa_5 n^{\frac{\gamma}{\eta}}) - 3\mathbb{P}(\mathcal{E}_2^c) \\ &\geq 1 - \exp(-t' \kappa_5 n^{\frac{\gamma}{\eta}}) - 3\exp(-\kappa_6 n^{1+\frac{\gamma(1-\eta)}{\eta}}) \\ &\geq 1 - 4\exp(-\min\{t'\kappa_5, \kappa_6\} \cdot n^{\frac{\gamma}{\eta}}) \end{aligned} \tag{A.78}$$

where we have used the previous lemma to bound $\mathbb{P}(\mathcal{E}_2^c)$, and also the assumption $\gamma \in (0,1)$ to conclude that $\frac{\gamma}{\eta} \leq 1+\frac{\gamma(1-\eta)}{\eta}$. Taking a union bound over $i = 1, \dots, n$, proves the claim. □

The last component of Lemma 2.2 is to prove the MSPE inequalities.

Proof of the MSPE inequalities

The proof of Theorem (2.3) shows that for any realization of X we have

$$\operatorname{mspe}(\widehat{\beta}_\varrho | X) := \tfrac{1}{n}\mathbb{E}\Big[\|X(\widehat{\beta}_\varrho - \beta)\|_2^2 \, \Big| \, X\Big] \lesssim n^{-\theta}\|\beta\|_2^2 + \tfrac{1}{n}\sum_{i=1}^{p\wedge n} \big(\tfrac{l_i}{l_i+\varrho}\big)^2, \tag{A.79}$$

where $l_i = \lambda_i(\frac{1}{n}X^\top X)$. Now observe that the second term on the right side matches the expression for $\text{tr}(M(X))$ given in line (A.46) by replacing ρ with ϱ and multiplying by a factor of $\frac{1}{n}$. Therefore, using Lemma A.8 and recalling $\varrho/n = n^{-\theta}$ shows that there are absolute constants $\kappa_1, \kappa_2, \kappa_3, \kappa_4 > 0$ such that for all large n,

$$\mathbb{P}\Big[\text{mspe}(\widehat{\beta}_\varrho|X) \geq \kappa_1\big(n^{-\theta} + n^{2(\theta-\eta)}\big)\Big] \leq \exp(-\kappa_2 n^{2(1-\eta)+\theta}), \quad \text{if } \ \eta \in (0, \tfrac{1}{2}). \tag{A.80}$$

and

$$\mathbb{P}\Big[\text{mspe}(\widehat{\beta}_\varrho|X) \geq \kappa_3\big(n^{-\theta} + n^{\frac{\theta}{\eta}-1}\big)\Big] \leq \exp(-\kappa_4 n^{1+\frac{\theta(1-\eta)}{\eta}}), \qquad \text{if } \ \eta > \tfrac{1}{2}. \tag{A.81}$$

In line (A.80), the bound $\text{mspe}(\widehat{\beta}_\varrho|X)$ is optimized when $n^{-\theta} \asymp n^{2(\theta-\eta)}$, which explains the choice $\theta = \frac{2\eta}{3}$. Similarly, in line (A.81), the bound is optimized when $n^{-\theta} \asymp n^{\frac{\theta}{\eta}-1}$, which explains the choice $\theta = \frac{\eta}{\eta+1}$. Substituting in these values θ yields the stated result. □

A.5 Background results

Results on matrices and convexity

Lemma A.10. *Let* $M_1 \in \mathbb{R}^{k_1 \times k_1}$ *and* $M_2 \in \mathbb{R}^{k_1 \times k_2}$. *Then,*

$$\|M_1 M_2\|_F \leq \|M_1\|_{\text{op}} \|M_2\|_F. \tag{A.82}$$

Proof. Observe that

$$\begin{aligned}
\|M_1M_2\|_F^2 &= \text{tr}(M_2^\top M_1^\top M_1 M_2)\\
&= \text{tr}((M_1^\top M_1)(M_2 M_2^\top))\\
&\leq \sum_{i=1}^{k_1} \lambda_i(M_1^\top M_1) \cdot \lambda_i(M_2 M_2^\top)\\
&\leq \|M_1\|_{\text{op}}^2 \sum_{i=1}^{k_1} \lambda_i(M_2 M_2^\top)\\
&= \|M_1\|_{\text{op}}^2 \|M_2\|_F^2,
\end{aligned} \tag{A.83}$$

where we have used von Neumann's trace inequality (also known as Fan's inequality) [BL00, p.10] in the third line. □

A result on convex trace functionals. In the following lemma, an interval of the real line refers to any set of the form (a, b),$(a, b]$,$[a, b)$, or $[a, b]$, where $-\infty \leq a \leq b \leq \infty$. We also define $\text{spec}(M)$ to be the set of eigenvalues of a square matrix M. The collection of symmetric matrices in $\mathbb{R}^{p\times p}$ is denoted by $\mathbb{S}^{p\times p}$. For a univariate function φ, the symbol $\text{tr}(\varphi(M))$ denotes $\sum_i \varphi(\lambda_i(M))$.

Lemma A.11. *Let $\mathcal{I} \subset \mathbb{R}$ be an interval, and let $\mathcal{M} \subset \mathbb{S}^{p\times p}$ be a convex set such that spec(M) $\subset \mathcal{I}$ for all $M \in \mathcal{M}$. Let $\varphi : \mathcal{I} \to \mathbb{R}$ be a convex function. Then, the functional*

$$M \mapsto \operatorname{tr}(\varphi(M)) \tag{A.84}$$

is convex on $\mathcal{M}$.

A proof may be found in the paper [Pet94, Proposition 2].

Operator Jensen inequality. A function $f : \mathbb{S}^{p\times p} \to \mathbb{S}^{p\times p}$ is said to be operator convex if for all $\lambda \in [0, 1]$, and all $A, B \in \mathbb{S}^{p\times p}$,

$$f(\lambda A + (1 - \lambda)B) \preceq \lambda f(A) + (1 - \lambda)f(B), \tag{A.85}$$

where $A \preceq B$ means that $B - A$ is positive semidefinite.

Lemma A.12 (Operator Jensen inequality)**.** *Suppose $f : \mathbb{S}^{p\times p} \to \mathbb{S}^{p\times p}$ is operator convex, and let A be a random $\mathbb{S}^{p\times p}$-valued matrix that is integrable. Then,*

$$f(\mathbb{E}[A]) \preceq \mathbb{E}[f(A)]. \tag{A.86}$$

Proof. It is enough to show that for all $x \in \mathbb{R}^p$,

$$x^\top f(\mathbb{E}[A])x \leq x^\top \mathbb{E}[f(A)]x. \tag{A.87}$$

For any fixed x, consider the function $g : \mathbb{S}^{p\times p} \to \mathbb{R}$ defined by $g(A) = x^\top f(A)x$. It is clear that g is a convex function in the usual sense, and so the ordinary version of Jensen's inequality implies $g(\mathbb{E}[A]) \leq \mathbb{E}[g(A)]$, which is the same as (A.87). □

Results on Gaussian vectors and matrices

The following lemma is standard and is often referred to as the Gaussian concentration inequality [BLM13].

Lemma A.13. *Let $Z \in \mathbb{R}^p$ be a standard Gaussian vector and let $f : \mathbb{R}^p \to \mathbb{R}$ be an L-Lipschitz function with respect to the ℓ_2 norm. Then for all $t > 0$,*

$$\mathbb{P}\Big(|f(Z) - \mathbb{E}[f(Z)]| \geq t\Big) \leq 2\exp\left(\tfrac{-t^2}{2L^2}\right). \tag{A.88}$$

Next, we give a formula for the expected square of a Wishart matrix.

Lemma A.14. *Let $X \in \mathbb{R}^{n\times p}$ have rows drawn i.i.d. from $N(0, \Sigma)$, and let $\widehat{\Sigma} = \frac{1}{n}X^\top X$. Then,*

$$\mathbb{E}[\widehat{\Sigma}^2] = (1 + \tfrac{1}{n})\Sigma^2 + \tfrac{\operatorname{tr}(\Sigma)}{n}\Sigma.$$

Proof. Write $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ where $X_i^\top \in \mathbb{R}^p$ is the ith row of X. If we put $M_i = X_i X_i^\top$, then

$$\widehat{\Sigma}^2 = \tfrac{1}{n^2}\textstyle\sum_{i=1}^n M_i^2 + \tfrac{1}{n^2}\sum_{i\neq j} M_i M_j.$$

Clearly, the M_i are independent with $\mathbb{E}[M_i] = \Sigma$ for all i, and so

$$\mathbb{E}[\widehat{\Sigma}^2] = \tfrac{1}{n}\mathbb{E}[M_1^2] + \tfrac{2}{n^2}\tbinom{n}{2}\Sigma^2.$$

It remains to compute $\mathbb{E}[M_1^2]$. Write $X_i = \Sigma^{1/2} Z_i$ where Z_i is a standard Gaussian vector in $\mathbb{R}^p$, and let $\Sigma^{1/2} = U\Lambda^{1/2}U^\top$ be a spectral decomposition of $\Sigma^{1/2}$ where $U \in \mathbb{R}^{p\times p}$ is orthogonal, and Λ is diagonal with $\Lambda_{ii} = \lambda_i(\Sigma)$. By the orthogonal invariance of the normal distribution, $X_i \overset{\mathcal{L}}{=} U\Lambda^{1/2} Z_i$, and so

$$\mathbb{E}[M_1^2] = \mathbb{E}[X_1 X_1^\top X_1 X_1^\top] = U\Lambda^{1/2}\mathbb{E}\Big[Z_1 Z_1^\top \Lambda Z_1 Z_1^\top\Big]\Lambda^{1/2}U^\top. \tag{A.89}$$

Define the matrix $\tilde{M}_1 := Z_1 Z_1^\top \Lambda Z_1 Z_1^\top$. It is straightforward to verify that $\mathbb{E}[\tilde{M}_1]$ is diagonal, and its jth diagonal entry is

$$\mathbb{E}[\tilde{M}_1]_{jj} = \operatorname{tr}(\Sigma) + 2\lambda_j.$$

Therefore,

$$\mathbb{E}[\tilde{M}_1] = \operatorname{tr}(\Sigma) I_{p\times p} + 2\Lambda,$$

and combining this with line (A.89) gives

$$\mathbb{E}[M_1^2] = \operatorname{tr}(\Sigma)\Sigma + 2\Sigma^2.$$

□

Signed s.v.d. The following lemma describes the factors of an s.v.d. of a standard Gaussian matrix $Z \in \mathbb{R}^{n\times p}$ with $n \geq p$. To make the statement of the lemma more concise, we define the term *signed s.v.d.* below. This is merely a particular form of the s.v.d. that ensures uniqueness. Specifically, if $Z \in \mathbb{R}^{n\times p}$ is a full rank matrix with $n \geq p$, then the signed s.v.d. of Z is given by

$$Z = HLG^\top, \tag{A.90}$$

where $H \in \mathbb{R}^{n\times p}$ has orthonormal columns, the matrix $L \in \mathbb{R}^{p\times p}$ is diagonal with $L_{11} \geq L_{22} \geq \cdots \geq L_{pp} > 0$, and $G \in \mathbb{R}^{p\times p}$ is orthogonal with its first row non-negative, i.e. $G_{1i} \geq 0$ for all $i = 1, \ldots, p$. It is a basic fact from linear algebra that the signed s.v.d. of any full rank matrix in $\mathbb{R}^{n\times p}$ exists and is unique [HJ09, Lemma 7.3.1]. This fact applies to Gaussian matrices, since they are full rank with probability 1.

Lemma A.15. *Suppose $n \geq p$, and let $Z \in \mathbb{R}^{n\times p}$ be a random matrix with entries drawn i.i.d. from $N(0,1)$. Let*

$$Z = HLG^\top, \tag{A.91}$$

be the unique signed s.v.d. of Z as defined above. Then, the matrices H, L and G are independent. Furthermore, if $H_i^\top \in \mathbb{R}^p$ denotes the ith row of H, then for each $i = 1, \ldots, n$, the marginal law of $H_i^\top$ is given by

$$H_i^\top \overset{\mathcal{L}}{=} \tfrac{1}{\|z\|_2} \cdot \Pi_p(z), \tag{A.92}$$

where $z \in \mathbb{R}^n$ is a standard Gaussian vector, and $\Pi_p(z)$ is the projection operator onto the first p coordinates, i.e. $\Pi_p(z) = (z_1, \ldots, z_p)$.

Proof. We first argue that H, L, and G are independent, and then derive the representation for $H_i^\top$ in the latter portion of the proof.

Due to the fact that the transformation $Z \mapsto (H, L, G)$ is invertible, it is possible to obtain the joint density of (H, L, G) from the density of Z by computing the matrix Jacobian of the factorization $Z = HLG$. (See the references [Mui82], [Mat97], and [ER05] for more background on Jacobians of matrix factorizations.) To speak in more detail about the joint density, let $\mathbb{V}^{n\times p}$ denote the Stiefel manifold of $n \times p$ matrices with orthonormal columns. Also, let $\mathbb{D}^{p\times p}$ denote the set of $p \times p$ diagonal matrices, and $\mathbb{O}^{p\times p}$ the set of orthogonal $p \times p$ matrices. The subset of $\mathbb{O}^{p\times p}$ with non-negative entries in the first row will be denoted by $\mathbb{O}_+^{n\times p}$.

Let $f_{H,L,G} : \mathbb{V}^{n\times p} \times \mathbb{D}^{p\times p} \times \mathbb{O}_+^{p\times p} \to [0, \infty)$ denote the joint density of (H, L, G), where the base measure is the product of the Haar distribution on $\mathbb{V}^{n\times p}$, Lebesgue measure on $\mathbb{D}^{p\times p}$, and the Haar distribution on $\mathbb{O}^{p\times p}$ restricted to $\mathbb{O}_+^{p\times p}$. (See the book [Chi03] for background on these measures). Then, it is known that $f_{H,L,G}$ factors according to

$$f_{H,L,G}(h, l, g) = f_H(h) \cdot f_L(l) \cdot f_G(g), \tag{A.93}$$

where $f_H : \mathbb{V}_+^{n\times p} \to [0, \infty)$ denotes the density of H with respect to the Haar measure on $\mathbb{V}^{n\times p}$, and similarly for f_L and f_G. A derivation of the factorization (A.93) via matrix Jacobians can be found in the paper [Jam54, Section 8]. This proves that H, L, and G are independent.

We now prove the representation (A.92). From line 8.10 in the paper [Jam54], it is known that f_H is the density of the Haar distribution on $\mathbb{V}^{n\times p}$. If $H \in \mathbb{R}^{n\times p}$ is a random matrix distributed according to f_H, then Theorem 2.2.1(ii) in the book [Chi03] implies that the rows $H_i^\top$ can be represented as

$$H_i^\top \overset{\mathcal{L}}{=} \Pi_p(J_i^\top) \tag{A.94}$$

where $J_i^\top \in \mathbb{R}^n$ is the ith row of a Haar-distributed random matrix in $\mathbb{O}^{n\times n}$. Furthermore, the rows $J_i^\top$ are uniformly distributed on the unit sphere on $\mathbb{R}^n$, and hence can be represented as $z/\|z\|_2$, where $z \in \mathbb{R}^n$ is a standard Gaussian vector. □

The leverage scores for Gaussian designs. Note that the hat matrix is invariant to the covariance structure of the design points. That is, if $X^\top = \Sigma^{1/2} Z^\top$ with $Z \in \mathbb{R}^{n\times p}$ being a standard Gaussian matrix, then

$$\begin{aligned} X(X^\top X)^{-1}X^\top &= Z\Sigma^{1/2}(\Sigma^{1/2}Z^\top Z^\top \Sigma^{1/2})^{-1}\Sigma^{1/2}Z^\top \\ &= Z(Z^\top Z)^{-1}Z^\top. \end{aligned} \tag{A.95}$$

Hence, there is no loss of generality in analyzing the hat matrix with $\Sigma = I_{p\times p}$.

Proposition A.1. *Suppose there are absolute constants $k_1, k_2 \in (0,1)$ such that $k_1 \leq p/n \leq k_2$ for all (n,p). Let $Z \in \mathbb{R}^{n\times p}$ be a random matrix with entries drawn i.i.d. from $N(0,1)$, and put $H = Z(Z^\top Z)^{-1}Z^\top$. Then as $(n,p) \to \infty$,*

$$\max_{1\leq i\leq n} |H_{ii} - \tfrac{p}{n}| \to 0 \quad \textit{almost surely.} \tag{A.96}$$

Proof. Let $Z = QR$ be a QR factorization where $Q \in \mathbb{R}^{n\times p}$ and $R \in \mathbb{R}^{p\times p}$. Since $Q^\top Q = I_{p\times p}$, we have

$$H = QR(R^\top Q^\top QR)^{-1}R^\top Q^\top = QQ^\top. \tag{A.97}$$

It is a basic fact that Q follows that Haar distribution on the Steifel manifold of $n\times p$ matrices with orthonormal columns. Furthermore, the proof of Lemma A.15 shows that the rows of Q can be represented in distribution as $\Pi_p(z)/\|z\|_2$ where $z \in \mathbb{R}^n$ is a standard Gaussian vector, and $\Pi_p(z) = (z_1, \dots, z_p)$. Hence, for each $i = 1, \dots, n$,

$$H_{ii} = e_i^\top He_i = e_i^\top QQ^\top e_i \overset{\mathcal{L}}{=} \tfrac{\|\Pi_p(z)\|_2^2}{\|z\|_2^2}. \tag{A.98}$$

It is clear that line (A.96) is implied by the two following limits

$$\max_{1\leq i\leq n} \tfrac{n}{p} H_{ii} \to 1 \quad \text{almost surely,} \tag{A.99}$$

and

$$\min_{1\leq i\leq n} \tfrac{n}{p} H_{ii} \to 1 \quad \text{almost surely.} \tag{A.100}$$

We will only show (A.99), since the proof of the second limit is essentially the same. With z as above, it is simple to verify the following concentration bound using lemma A.2,

$$\mathbb{P}\Big(a_n \leq \tfrac{\|\Pi_p(z)\|_2^2/p}{\|z\|_2^2/n} - 1 \leq b_n\Big) \geq 1 - 4\exp(-p^{1/4}), \tag{A.101}$$

where a_n and b_n are numerical sequences that tend to 0.

Consequently, the union bound gives

$$\mathbb{P}\Big(a_n \leq \max_{1\leq i\leq n} H_{ii} - 1 \leq b_n\Big) \geq 1 - 4n \cdot \exp(-p^{1/4}) \tag{A.102}$$

Since p/n is bounded strictly between 0 and 1, the series $\sum_{n\geq 1} n\exp(-p^{1/4})$ is finite, and then the Borel-Cantelli lemma implies (A.99). □

Appendix B

Proofs for Chapter 3

B.1 Proofs for Section 3.1

The relation (3.12)

Proof. It is enough to show that ψ^{-1} has a strictly negative derivative at all positive values of $\tau = \tau(x)$. Direct calculation shows that the condition $(\psi_p^{-1})'(\tau(x)) < 0$ is equivalent to the inequality

$$\frac{\sum_{i=1}^{p} i^{-\tau(x)}}{\sum_{i=1}^{p} \log(i) i^{-\tau(x)}} \cdot \frac{\sum_{i=1}^{p} \log(i) i^{-2\tau(x)}}{\sum_{i=1}^{p} \log(i) i^{-\tau(x)}} < \frac{\sum_{i=1}^{p} i^{-2\tau(x)}}{\sum_{i=1}^{p} \log(i) i^{-\tau(x)}}. \tag{B.1}$$

To prove this inequality, consider the probability mass function

$$\pi(j) := \frac{\log(j) j^{-\tau(x)}}{\sum_{i=1}^{p} \log(i) i^{-\tau(x)}}$$

on the set $\{1, \dots, p\}$. By interpreting the sums in line (B.1) in terms of expectations with respect to a discrete random variable $J \sim \pi$, the inequality is equivalent to

$$\mathbb{E}\big[\tfrac{1}{\log J}\big]\mathbb{E}\big[J^{-\tau(x)}\big] < \mathbb{E}\big[\tfrac{J^{-\tau(x)}}{\log J}\big]. \tag{B.2}$$

Since the functions $j \mapsto j^{-\tau(x)}$ and $j \mapsto 1/\log(j)$ are strictly decreasing on $\{1, \dots, p\}$, it follows from the "association inequality" ([BLM13], Theorem 2.14) that line (B.2) is true. □

B.2 Proofs for Section 3.2

Proposition 3.1

Proof. Define the positive number $t := \frac{n}{\log(\frac{pe}{n})}$, and choose $T = \lceil t \rceil$ in Theorem 3.1. We first verify that $T \log(\frac{pe}{T}) \leq 3n$ for every n. Observe that

$$\begin{aligned} T\log(\tfrac{pe}{T}) &\leq (t+1)\log(\tfrac{pe}{t}) \\ &= \tfrac{n}{\log(\frac{pe}{n})} \log\Big(\tfrac{pe}{n} \cdot \log(\tfrac{pe}{n})\Big) + \log\Big(\tfrac{pe}{n} \cdot \log(\tfrac{pe}{n})\Big). \end{aligned} \tag{B.3}$$

Let $r = p/n$ and recall that we assume $n \leq p$. Simple calculus shows that for all $r \geq 1$, the quantity $\log\big(re \cdot \log(re)\big)$ is at most $1.4\log(re)$, and so

$$\begin{aligned} T\log(\tfrac{pe}{T}) &\leq \tfrac{n}{\log(\frac{pe}{n})} \cdot 1.4\log(\tfrac{pe}{n}) + 1.4\log(\tfrac{pe}{n}) \\ &\leq 1.4n + 1.4n \\ &\leq 3n \end{aligned} \tag{B.4}$$

with the second step following from the assumption $\log(\frac{pe}{n}) \leq n$. Consequently, the condition (3.15) of Theorem 3.1 is satisfied for every n under this choice of T, and we conclude that there is an absolute constant $c_1 > 0$ such that the bound (3.16) holds with probability at least $1 - 2\exp(-c_1 n)$. To finish the argument, observe that

$$\tfrac{1}{\sqrt{T}}\|x - x_{|T}\|_1 \leq \tfrac{1}{\sqrt{t}}\|x\|_1 = \tfrac{1}{\sqrt{n}}\sqrt{\|x\|_1^2 \log(\tfrac{pe}{n})}. \tag{B.5}$$

Dividing the inequality (3.16) through by $\|x\|_2$ leads to

$$\tfrac{\|x-\widehat{x}\|_2}{\|x\|_2} \leq c_2\tfrac{\sigma\epsilon_0}{\|x\|_2} + \tfrac{c_3}{\sqrt{n}}\sqrt{\tfrac{\|x\|_1^2}{\|x\|_2^2}\log(\tfrac{pe}{n})},$$

and the proof is complete. □

Proposition 3.2

Proof. Let $\mathbb{B}_1(1) \subset \mathbb{R}^p$ be the ℓ_1 ball of radius 1, and let $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}^p$ be a homogeneous recovery algorithm. Also define the number $\eta := \frac{1}{2}c_1\sqrt{\frac{\log(pe/n)}{n}}$ where $c_1 > 0$ is an absolute constant to be defined below. Clearly, for any such η, we can find a point $\tilde{x} \in \mathbb{B}_1(1)$ satisfying

$$\|\tilde{x} - \mathcal{R}(A\tilde{x})\|_2 \geq \sup_{x \in \mathbb{B}_1(1)} \|x - \mathcal{R}(Ax)\|_2 - \eta. \tag{B.6}$$

Furthermore, we may choose such a point $\tilde{x}$ to satisfy $\|\tilde{x}\|_1 = 1$. (Note that if $\|\tilde{x}\|_1 < 1$, then we can use the homogeneity of $\mathcal{R}$ to replace $\tilde{x}$ with the ℓ_1 unit vector $\tilde{x}/\|\tilde{x}\|_1 \in \mathbb{B}_1(1)$ and obtain an even larger value on the left side.) The next step of the argument is to further lower bound the right side in terms of minimax error, leading to

$$\|\tilde{x} - \mathcal{R}(A\tilde{x})\|_2 \geq \inf_{A \in \mathbb{R}^{n\times p}} \inf_{R\colon \mathbb{R}^n \to \mathbb{R}^p} \sup_{x \in \mathbb{B}_1(1)} \|x - R(Ax)\|_2 - \eta, \tag{B.7}$$

where the infima are over all sensing matrices A and all recovery algorithms R (possibly non-homogenous). It is known from the theory of Gelfand widths that there is an absolute constant $c_1 > 0$, such that the minimax ℓ_2 error over $\mathbb{B}_1(1)$ is lower-bounded by

$$\inf_{A \in \mathbb{R}^{n\times p}} \inf_{R\colon \mathbb{R}^n \to \mathbb{R}^p} \sup_{x \in \mathbb{B}_1(1)} \|x - R(Ax)\|_2 \geq c_1\sqrt{\tfrac{\log(pe/n)}{n}}. \tag{B.8}$$

(See [Can06, Section 3.5], as well as [Kas77] [GG84].) Using our choice of η, as well as the fact that $\|\tilde{x}\|_1 = 1$, we obtain

$$\|\tilde{x} - \mathcal{R}(A\tilde{x})\|_2 \geq \tfrac{1}{2}c_1\sqrt{\tfrac{\|\tilde{x}\|_1^2 \cdot \log(pe/n)}{n}}. \tag{B.9}$$

Dividing both sides by $\|\tilde{x}\|_2$ completes the proof. □

B.3 Proofs for Section 3.3

Theorem 3.2 – Uniform CLT for ℓ_q norm estimator

The proof of Theorem 3.2 consists of two parts. First, we prove a uniform CLT for a re-scaled version of $\widehat{\Psi}_n(t)$, which is given below in Lemma B.1. Second, we extend this limit to the statistic $\widehat{\nu}_q(t)$ by way of the functional delta method, which is described at the end of the section.

Remark on the subscript of n_q. For ease of notation, we will generally drop the subscript from n_q in the remainder of the appendix, as it will not cause confusion.

Weak convergence of the empirical characteristic function. To introduce a few pieces of notation, let $c \in [-b, b]$ for some fixed $b > 0$, and define the re-scaled empirical characteristic function,

$$\widehat{\psi}_n(c) := \frac{1}{n}\sum_{i=1}^{n} e^{\sqrt{-1}\frac{cy_i}{\gamma_q\|x\|_q}}, \tag{B.10}$$

which is obtained from the re-scaled observations $\frac{y_i}{\gamma\|x\|_q} \overset{d}{=} S_i + \rho_q\epsilon_i$, where $S_i \sim \text{stable}_q(1)$, and $\epsilon_i \sim F_0$. The relation between $\widehat{\Psi}_n$ and $\widehat{\psi}_n$ is given by

$$\widehat{\Psi}_n(t) = \widehat{\psi}_n(\gamma_q t\|x\|_q). \tag{B.11}$$

The re-scaled population characteristic function is

$$\psi_n(c) := \exp(-|c|^q)\varphi_0(\rho_q c), \tag{B.12}$$

which converges to the function

$$\psi(c) := \exp(-|c|^q)\varphi_0(\bar{\rho}_q c), \tag{B.13}$$

as $\rho_q \to \bar{\rho}_q$. Lastly, define the normalized process

$$\chi_n(c) := \sqrt{n}\Big(\widehat{\psi}_n(c) - \psi(c)\Big), \tag{B.14}$$

and let $\mathscr{C}([-b, b]; \mathbb{C})$ be the space of continuous complex-valued functions on $[-b, b]$ equipped with the sup-norm.

Lemma B.1. *Fix any $b > 0$. Under the assumptions of Theorem 3.2, the random function χ_n satisfies the limit*

$$\chi_n(c) \xrightarrow{w} \chi_\infty(c) \quad \textit{in} \quad \mathscr{C}([-b,b];\mathbb{C}), \tag{B.15}$$

where χ_∞ is a centered Gaussian process whose marginals satisfy $\mathrm{Re}(\chi_\infty(c)) \sim N(0, \omega(c, \bar\rho_q))$, and

$$\omega(c, \bar\rho_q) = \tfrac{1}{2} + \tfrac{1}{2}\exp(-2^q|c|^q)\varphi_0(2\bar\rho_q c) - \exp(-2|c|^q)\varphi_0^2(\bar\rho_q c). \tag{B.16}$$

Proof. It is important to notice that $\widehat{\psi}_n(c)$ is not the empirical characteristic function associated with n samples from the distribution of ψ (since $\rho_q \neq \bar\rho_q$). The more natural process to work with is

$$\tilde\chi_n(c) := \sqrt{n}\big(\tilde\psi_n(c) - \psi(c)\big), \tag{B.17}$$

where $\tilde\psi_n(c) = \frac{1}{n}\sum_{i=1}^n \exp(\sqrt{-1}cy_i^\circ)$ and $y_i^\circ = S_i + \bar\rho_q\epsilon_i$. (In other words, $\tilde\psi_n$ is the empirical characteristic function associated with ψ.) As a first step in the proof, we show that the difference between χ_n and $\tilde\chi_n$ is negligible in a uniform sense, i.e.

$$\sup_{c\in[-b,b]} |\chi_n(c) - \tilde\chi_n(c)| = o(1) \text{ a.s.} \tag{B.18}$$

To see this, observe that for $c \in [-b, b]$,

$$\begin{aligned}
|\chi_n(c) - \tilde\chi_n(c)| &= \sqrt{n}|\widehat{\psi}_n(c) - \tilde\psi_n(c)| && \text{(B.19)}\\
&= \tfrac{1}{\sqrt{n}}\Big|\textstyle\sum_{i=1}^n e^{\sqrt{-1}c(S_i+\rho_q\epsilon_i)} - e^{\sqrt{-1}c(S_i+\bar\rho_q\epsilon_i)}\Big| && \text{(B.20)}\\
&= \tfrac{1}{\sqrt{n}}\Big|\textstyle\sum_{i=1}^n e^{\sqrt{-1}c(S_i+\rho_q\epsilon_i)}\big(1 - e^{\sqrt{-1}c(\bar\rho_q-\rho_q)\epsilon_i}\big)\Big| && \text{(B.21)}\\
&\leq \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n \big|1 - e^{\sqrt{-1}c(\bar\rho_q-\rho_q)\epsilon_i}\big| && \text{(B.22)}\\
&\leq \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n |c(\bar\rho_n - \rho_q)\epsilon_i| && \text{(B.23)}\\
&\leq \sqrt{n}|\rho_q - \bar\rho_q| \cdot \tfrac{b}{n}\textstyle\sum_{i=1}^n |\epsilon_i|, && \text{(B.24)}
\end{aligned}$$

where the last bound does not depend on c, and tends to 0 almost surely. Here we are using the assumption that $\mathbb{E}|\epsilon_1| < \infty$ and assumption **A3.3** that $\rho_q = \bar\rho_q + o(1/\sqrt{n})$. Now that line (B.18) has been verified, it remains (by the functional version of Slutsky's Lemma [VW96, p.32]) to prove

$$\tilde\chi_n \xrightarrow{w} \chi_\infty \quad \text{in} \quad \mathscr{C}([-b,b];\mathbb{C}), \tag{B.25}$$

and that the limiting process χ_∞ has the stated variance formula. We first show that this limit holds, and then derive the variance formula at the end of the proof. (Note that it is clear that the limiting process must be Gaussian due to the finite-dimensional CLT.)

By a result of Marcus [Mar81, Theorem 1], it is known that the uniform CLT for empirical characteristic functions (B.25) holds as long as the limiting process χ_∞ has continuous sample paths (almost surely).[1] To show that the sample paths of χ_∞ are continuous, we employ as sufficient condition derived by Csörgo [Cso81a]. Let F_q denote the distribution function of the random variable $y_i^\circ = S_i + \bar{\rho}_q \epsilon_i$ described earlier. Also let $\delta > 0$ and define the function

$$g_\delta^+(u) = \begin{cases} \log(|u|) \cdot \log(\log(|u|))^{2+\delta} & \text{if } \; |u| \geq \exp(1) \\ 0 & \text{if } \; |u| < \exp(1). \end{cases} \tag{B.26}$$

At line 1.17 of the paper [Cso81a], it is argued that if

$$\int_{-\infty}^{\infty} g_\delta^+(|u|) dF_q(u) < \infty, \tag{B.27}$$

then χ_∞ has continuous sample paths. Next, note that for any $\delta, \delta' > 0$, we have

$$g_\delta^+(|u|) = \mathcal{O}\Big(|u|^{\delta'}\Big) \qquad \text{as } |u| \to \infty. \tag{B.28}$$

Hence, χ_∞ has continuous sample paths as long as F_q has a fractional moment. To see that this is true, recall the basic fact that if $S_i \sim \text{stable}_q(1)$ then $\mathbb{E}[|S_i|^{q'}] < \infty$ for any $q' \in (0, q)$. Also, we assume that $\mathbb{E}[|\epsilon_i|] < \infty$, and so it follows that for any $q \in (0, 2]$, the distribution F_q has a fractional moment, which proves the limit (B.25).

Finally, we compute the variance of the marginal distributions $\text{Re}(\chi_\infty(c))$. By the ordinary central limit theorem, we only need to calculate the variance of $\text{Re}(\exp(\sqrt{-1}y_1^\circ)$. The first moment is given by

$$\mathbb{E}[\text{Re}(\exp(\sqrt{-1}cy_1^\circ)] = \text{Re}\,\mathbb{E}[\exp(\sqrt{-1}cy_1^\circ)] = \exp(-|c|^q)\varphi_0(\bar{\rho}_q c).$$

The second moment is given by

$$\begin{aligned} \mathbb{E}[(\text{Re}(\exp(\sqrt{-1}cy_1^\circ))^2] &= \mathbb{E}[\cos^2(cy_1^\circ)] & \text{(B.29)} \\ &= \mathbb{E}[\tfrac{1}{2} + \tfrac{1}{2}\cos(2cy_1^\circ)] & \text{(B.30)} \\ &= \tfrac{1}{2} + \tfrac{1}{2}\text{Re}\,\mathbb{E}[\exp(\sqrt{-1} \cdot 2cy_1^\circ)] & \text{(B.31)} \\ &= \tfrac{1}{2} + \tfrac{1}{2}\exp(-2^q|c|^q)\varphi_0(2\bar{\rho}_q c). & \text{(B.32)} \end{aligned}$$

This completes the proof of Lemma B.1.

□

Applying the functional delta method. We now complete the proof of Theorem 3.2 by applying the functional delta method to Lemma B.1 (with a suitable map ϕ to be defined below). In the following, $\mathscr{C}(\mathcal{I})$ denotes the space of continuous real-valued functions on an interval $\mathcal{I}$, and $\ell^\infty(\mathcal{I})$ denotes the space of bounded real-valued functions on $\mathcal{I}$. Both are

[1] The paper [Mar81] only states the result when $b = 1/2$, but it holds for any $b > 0$. See the paper [Cso81b, Theorem 3.1].

equipped with the sup-norm.

Proof of Theorem 3.2. Since $\varphi_0(\bar{\rho}_q c_0) \neq 0$, there is some $\delta_0 \in (0, |c_0|)$ such that over the interval $c \in \mathcal{I} := [c_0 - \delta_0, c_0 + \delta_0]$, the value $\varphi_0(\bar{\rho}_q c)$ is bounded away from 0. Define the function $f_0 \in \mathscr{C}(\mathcal{I})$ by

$$f_0(c) = \exp(-|c|^q),$$

and let $\mathscr{N}(f_0; \varepsilon) \subset \mathscr{C}(\mathcal{I})$ be a fixed ε-neighborhood of f_0 in the sup-norm, such that all functions in the neighborhood are bounded away from 0. Consider the map $\phi : \mathscr{C}(\mathcal{I}) \to \ell^\infty(\mathcal{I})$ defined according to

$$\phi(f)(c) = \begin{cases} -\frac{1}{|c|^q}\log(f(c)) & \text{if } f \in \mathscr{N}(f_0;\varepsilon) \\ 1 & \text{if } f \notin \mathscr{N}(f_0;\varepsilon). \end{cases} \tag{B.33}$$

The importance of ϕ is that it can be related to $\widehat{\nu}_q(\widehat{t})/\|x\|_q^q$ in the following way. First, let $\widehat{c} = \widehat{t}\gamma_q\|x\|_q$ and observe that the definition of $\widehat{\psi}_n$ gives

$$\sqrt{n}\big(\tfrac{\widehat{\nu}_q(\widehat{t})}{\|x\|_q^q} - 1\big) = \sqrt{n}\Big(- \tfrac{1}{|\widehat{c}|^q}\mathrm{Log}_+\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\widehat{c})}{\varphi_0(\rho_q\widehat{c})}\big)\big) + \tfrac{1}{|\widehat{c}|^q}\mathrm{Log}_+\big(\exp(-|\widehat{c}|^q)\big)\Big). \tag{B.34}$$

Next, let $\Pi(\widehat{c})$ be the point in the interval $\mathcal{I}$ that is nearest to $\widehat{c}$,[2] and define the quantities Δ_n and Δ_n' according to

$$\sqrt{n}\big(\tfrac{\widehat{\nu}_q(\widehat{t})}{\|x\|_q^q} - 1\big) = \sqrt{n}\Big(- \tfrac{1}{|\Pi(\widehat{c})|^q}\mathrm{Log}_+\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\Pi(\widehat{c}))}{\varphi_0(\rho_q\Pi(\widehat{c}))}\big)\big) + \tfrac{1}{|\Pi(\widehat{c})|^q}\mathrm{Log}_+\big(\exp(-|\Pi(\widehat{c})|^q)\big)\Big) + \Delta_n \tag{B.35}$$

$$= \sqrt{n}\Big(\phi\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\cdot)}{\varphi_0(\rho_q\cdot)}\big)\big)(\Pi(\widehat{c})) - \phi\big(f_0\big)(\Pi(\widehat{c}))\Big) + \Delta_n' + \Delta_n. \tag{B.36}$$

We now argue that both of the terms Δ_n and Δ_n' are asymptotically negligible. As a result, we may complete the proof of Theorem 3.2 by showing that the first term in line (B.36) has the desired Gaussian limit — which is the purpose of the functional delta method.

To see that $\Delta_n \to_P 0$, first recall that $\widehat{c} \to_P c_0 \in \mathcal{I}$ by assumption. Consequently, along any subsequence, there is a further subsequence on which $\widehat{c}$ and $\Pi(\widehat{c})$ eventually agree with probability 1. In turn, if g_n is a generic sequence of functions, then eventually $g_n(\widehat{c}) - g_n(\Pi(\widehat{c})) = 0$ along subsequences (with probability 1). Said differently, this means $g_n(\widehat{c}) - g_n(\Pi(\widehat{c})) \to_P 0$, and this implies $\Delta_n \to_P 0$ since Δ_n can be expressed in the form $g_n(\widehat{c}) - g_n(\Pi(\widehat{c}))$.

Next, to see that $\Delta_n' \to_P 0$, notice that as soon as $\widehat{\psi}_n(\cdot)/\varphi_0(\rho_q\cdot)$ lies in the neighborhood $\mathscr{N}(f_0;\varepsilon)$, it follows from the definition of ϕ that $\Delta_n' = 0$. Also, this is guaranteed to happen

[2] The purpose of introducing $\Pi(\widehat{c})$ is that it always lies in the interval $\mathcal{I}$, and hence allows us to work entirely on $\mathcal{I}$.

with probability 1 for large enough n, because the function $\widehat{\psi}_n(\cdot)/\varphi_0(\rho_q\cdot)$ converges uniformly to f_0 on the interval $\mathcal{I}$ with probability 1.[3] Hence, $\Delta_n \to 0$ almost surely, but we only need $\Delta_n' = o_P(1)$ in the remainder of the proof.

We have now taken care of most of the preparations needed to apply the functional delta method. Multiplying the limit in Lemma B.1 through by $1/\varphi_0(\rho_q\cdot)$ and using the functional version of Slutsky's Lemma [VW96, p.32], it follows that

$$\sqrt{n}\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\cdot)}{\varphi_0(\rho_q\cdot)}\big) - \exp(-|\cdot|^q)\big) \xrightarrow{w} \tilde{z}(\cdot) \quad \text{in} \quad \mathscr{C}(\mathcal{I}), \tag{B.37}$$

where $\tilde{z}(\cdot)$ is a centered Gaussian process with continuous sample paths, and the marginals $\tilde{z}(c)$ have variance equal to

$$\tfrac{1}{2}\tfrac{1}{\varphi_0^2(\bar{\rho}_q c)} + \tfrac{1}{2}\exp(-2^q|c|^q)\tfrac{\varphi_0(2\bar{\rho}_q c)}{\varphi_0^2(\bar{\rho}_q c)} - \exp(-2|c|^q). \tag{B.38}$$

It is straightforward to verify that ϕ is Hadamard differentiable at f_0 and the Hadamard derivative ϕ_{f_0}' is the linear map that multiplies by $-\frac{\exp|\cdot|^q}{|\cdot|^q}$. (See Lemmas 3.9.3 and 3.9.25 in [VW96].) Consequently, the functional delta method [VW96, Theorem 3.9.4] applied to line (B.37) with the map ϕ gives

$$\sqrt{n}\big(\phi\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\cdot)}{\varphi_0(\rho_q\cdot)}\big)\big) - \phi\big(\exp(-|\cdot|^q)\big)\big) \xrightarrow{w} \phi_{f_0}'(\tilde{z})(\cdot) \quad \text{in} \quad \ell^\infty(\mathcal{I}) \tag{B.39}$$

$$= -\tfrac{\exp(|\cdot|^q)}{|\cdot|^q}\tilde{z}(\cdot) \tag{B.40}$$

$$=: z(\cdot). \tag{B.41}$$

It is clear that $z(\cdot)$, defined in the previous line, is a centered Gaussian process on $\mathcal{I}$, since $\tilde{z}(\cdot)$ is. Combining lines (B.38) and (B.41) shows that the marginals $z(c)$ are given by

$$z(c) \sim N(0, v(c, \bar{\rho}_q)),$$

where

$$v_q(c, \bar{\rho}_q) = \tfrac{1}{|c|^{2q}}\Big(\tfrac{1}{2}\tfrac{1}{\varphi_0(\bar{\rho}_q c)^2}\exp(2|c|^q) + \tfrac{1}{2}\tfrac{\varphi_0(2\bar{\rho}_q c)}{\varphi_0(\bar{\rho}_q c)^2}\exp((2-2^q)|c|^q) - 1\Big). \tag{B.42}$$

The final step of the proof essentially involves plugging $\Pi(\widehat{c})$ into the limit (B.39) and using Slutsky's Lemma. Since $\Pi(\widehat{c}) \to_P c_0$, and $z(\cdot)$ takes values in the separable space $\mathscr{C}(\mathcal{I})$ almost surely[4], the functional version of Slutsky's Lemma [VW96, p.32] gives the following convergence of pairs

$$\Big(\sqrt{n}\big(\phi\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\cdot)}{\varphi_0(\rho_q\cdot)}\big)\big) - \phi\big(f_0\big)\big),\ \Pi(\widehat{c})\ \Big) \xrightarrow{w} \big(z(\cdot), c_0\big) \quad \text{in} \quad \ell^\infty(\mathcal{I}) \times \mathcal{I}. \tag{B.43}$$

[3] The fact that $\widehat{\psi}_n(\cdot)$ converges uniformly to $\psi(\cdot)$ on $\mathcal{I}$ with probability 1 follows essentially from the Glivenko-Cantelli Theorem [Cso81a, Equation 1.2]. To see that $1/\varphi_0(\rho_q\cdot)$ converges uniformly to $1/\varphi_0(\bar{\rho}_q\cdot)$ on $\mathcal{I}$ as $\rho_q \to \bar{\rho}_q$, note that since $\mathbb{E}|\epsilon_1| < \infty$, the characteristic function φ_0 is Lipschitz on any compact interval.

[4] This follows from the fact that $\tilde{z}(\cdot)$ has continuous sample paths almost surely.

To finish, note that the evaluation map $\ell^\infty(\mathcal{I})\times\mathcal{I} \to \mathbb{R}$ defined by $(f, c) \mapsto f(c)$, is continuous. The continuous mapping theorem [VW96, Theorem 1.3.6] then gives

$$\sqrt{n}\Big(\phi\big(\mathrm{Re}\big(\tfrac{\widehat{\psi}_n(\cdot)}{\varphi_0(\rho_q\cdot)}\big)\big)(\Pi(\widehat{c})) - \phi\big(f_0\big)(\Pi(\widehat{c}))\Big) \xrightarrow{w} z(c_0), \tag{B.44}$$

which is the desired conclusion. □

Lemma 3.2 – Extending the variance function

Proof. It is simple to verify that $v_q(\cdot,\cdot)$ is continuous at any pair (c_0, ρ_0) for which $c_0 \neq 0$ and $\varphi_0(\rho_0 c_0) \neq 0$, and hence $\tilde{v}_q$ inherits continuity at those pairs. To show that $\tilde{v}_q$ is continuous elsewhere, it is necessary to handle two cases.

- First, we show that for any $q \in (0, 2]$, if (c_0, ρ_0) is a pair such that $\varphi_0(\rho_0 c_0) = 0$ and $c_0 \neq 0$, then $v_q(c_j, \rho_j) \to \infty$ for any sequence satisfying $(c_j, \rho_j) \to (c_0, \rho_0)$ with $\varphi_0(\rho_j c_j) \neq 0$ and $c_j \neq 0$ for all j. (Note that $v_q(\cdot,\cdot)$ is defined in a deleted neighborhood of (c_0, ρ_0) due to the assumption that the roots of φ_0 are isolated.)
- Second, we show that for any $q \in (0, 2)$, if $c_0 = 0$, then $v_q(c_j, \rho_j) \to \infty$ for any sequence $(c_j, \rho_j) \to (0, \rho_0)$, where $\rho_0 \geq 0$ is arbitrary.

To handle the first case where $c_0 \neq 0$, we derive a lower bound on $v_q(c, \rho)$ for all $q \in (0, 2]$ and all pairs where v_q is defined. Recall the formula

$$v_q(c, \rho) = \tfrac{1}{|c|^{2q}}\Big(\tfrac{1}{2}\tfrac{1}{\varphi_0(\rho c)^2}\exp(2|c|^q) + \tfrac{1}{2}\tfrac{\varphi_0(2\rho|c|)}{\varphi_0(\rho|c|)^2}\exp((2-2^q)|c|^q) - 1\Big). \tag{B.45}$$

The lower bound is obtained by manipulating the factor $\frac{1}{2}\frac{\varphi_0(2\rho c)}{\varphi_0(\rho c)^2}$. Consider the following instance of Jensen's inequality, followed by a trigonometric identity,

$$\begin{aligned}
\varphi_0(\rho c)^2 &= (\mathbb{E}[\cos(\rho c\epsilon_1)])^2 &\text{(B.46)}\\
&\leq \mathbb{E}[\cos^2(\rho\epsilon_1)] &\text{(B.47)}\\
&= \tfrac{1}{2} + \tfrac{1}{2}\mathbb{E}[\cos(2\rho c\epsilon_1)] &\text{(B.48)}\\
&= \tfrac{1}{2} + \tfrac{1}{2}\varphi_0(2\rho c), &\text{(B.49)}
\end{aligned}$$

which gives $\frac{1}{2}\frac{\varphi_0(2\rho c)}{\varphi_0(\rho)^2} \geq 1 - \frac{1}{2}\frac{1}{\varphi_0(\rho c)^2}$. Letting $\kappa_q := 2 - 2^q$ we have the lower bound

$$v_q(c, \rho) \geq \tfrac{1}{|c|^{2q}}\Big(\tfrac{1}{2}\tfrac{1}{\varphi_0(\rho|c|)^2}\Big(\exp(2|c|^q) - \exp(\kappa_q|c|^q)\Big) + \exp(\kappa_q|c|^q) - 1\Big), \tag{B.50}$$

which holds for all (c, ρ) where $v_q(c, \rho)$ is defined. Since the quantity

$$\big(\exp(2c_0^q) - \exp(\kappa_q c_0^q)\big)$$

is positive for all $q \in (0, 2]$ and $c_0 \neq 0$, it follows that $v_q(c_j, \rho_j) \to \infty$ as $\varphi_0(\rho_j c_j) \to \varphi_0(\rho_0 c_0) = 0$.

Next, we consider the second case where $(c_j, \rho_j) \to (c_0, \rho_0)$ with $c_0 = 0$ and $\rho_0 \geq 0$. Due to the fact that all characteristic functions satisfy $|\varphi_0| \leq 1$, and the fact that $\big(\exp(2|c|^q) - \exp(\kappa_q |c|^q)\big)$ is positive when $c \neq 0$, the previous lower bound gives

$$v_q(c, \rho) \geq \tfrac{1}{|c|^{2q}} \Big(\tfrac{1}{2} \exp(2|c|^q) + \tfrac{1}{2} \exp(\kappa_q |c|^q) - 1 \Big), \tag{B.51}$$

which again holds for all (c, ρ) where v_q is defined. Note that this bound does not depend on ρ. A simple calculation involving L'Hospital's rule shows that whenever $q \in (0, 2)$, the lower bound tends to ∞ as $c \to 0+$.

To finish the proof, we must show that for any $\rho \geq 0$, the function $v_q(\cdot, \rho)$ attains its minimum on the set $[\varepsilon_q, \infty)$. This is simple because the lower bound (B.51) tends to ∞ as $c \to \infty$. □

B.4 Proofs for Section 3.4

Proposition 3.3 – consistency of pilot estimator.

Proof. We first show that there is a positive constant c_1 such that $\widehat{t}_{\text{initial}} \gamma_q \|x\|_q \to_P c_1$. Note that for each $i = 1, \ldots, n$, we have $\frac{y_i}{\gamma_q \|x\|_q} \sim S_i + \frac{\sigma}{\gamma_q \|x\|_q} \epsilon_i$, where the S_i are i.i.d. samples from stable$_q(1)$. Let F_n denote the distribution function of the random variable $|S_1 + \frac{\sigma}{\gamma_q \|x\|_q} \epsilon_1|$ and let $\mathbb{F}_n$ be the empirical distribution function obtained from n samples from F_n. Then,

$$\begin{aligned} \tfrac{\widehat{m}_q}{\gamma_q \|x\|_q} &= \tfrac{1}{\gamma_q \|x\|_q} \cdot \text{med}(|y_1|, \ldots, |y_n|) && \text{(B.52)} \\ &= \text{med}\big(\tfrac{|y_1|}{\gamma_q \|x\|_q}, \ldots, \tfrac{|y_n|}{\gamma_q \|x\|_q}\big) && \text{(B.53)} \\ &= \text{med}(\mathbb{F}_n). && \text{(B.54)} \end{aligned}$$

Note also that $F_n \xrightarrow{w} F$, where F is the distribution function of the variable $|S_1 + \bar{\rho}_q \epsilon_1|$. Consequently, it follows from a standard argument given in Section B.6 of Appendix B that $\text{med}(\mathbb{F}_n) \to_P \text{med}(F)$, and then

$$\tfrac{\widehat{m}_q}{\gamma_q \|x\|_q} \to_P \text{med}(F) =: 1/c_1. \tag{B.55}$$

Altogether, we have verified that $\widehat{t}_{\text{initial}} = 1/\widehat{m}_q$ satisfies $\widehat{t}_{\text{initial}} \gamma_q \|x\|_q \to_P c_1$. Combining the previous limit with line (3.45) and Theorem 3.2 then proves line (3.51), which in turn implies line (3.52). □

Proposition 3.4 – consistency of $c^\star(\widehat{\rho}_q)$.

Proof. As described in Section 3.3, we use ξ as a implicit index in our asymptotics, and recall that by assumption **A3.3**, we have $\rho_q = \rho_q(\xi) \to \bar{\rho}_q$ as $\xi \to \infty$. As a preliminary step, we will show $c^\star(\rho_q(\xi))$ is a bounded sequence. Namely, we will show there is a fixed compact interval $[\varepsilon_q, c_{\max}]$ such that for all large ξ,

$$c^\star(\rho_q(\xi)) \in [\varepsilon_q, c_{\max}]. \tag{B.56}$$

To show this, let $\ell_q(c)$ denote the right hand side of the bound (B.51), which satisfies

$$\ell_q(c) \leq \tilde{v}_q(c, \rho) \tag{B.57}$$

for all $q \in (0, 2]$, all ρ, and all $c > 0$. Also let $\bar{v}$ be any number satisfying

$$\tilde{v}_q(c^\star(\bar{\rho}_q), \bar{\rho}_q) < \bar{v}. \tag{B.58}$$

Since $\tilde{v}_q(\cdot, \cdot)$ is continuous, it follows that as $\xi \to \infty$,

$$\tilde{v}_q(c^\star(\bar{\rho}_q), \rho_q(\xi)) \to \tilde{v}_q(c^\star(\bar{\rho}_q), \bar{\rho}_q) \tag{B.59}$$

and so line (B.58) forces us to conclude that

$$\tilde{v}_q(c^\star(\bar{\rho}_q), \rho_q(\xi)) < \bar{v} \text{ for large } \xi. \tag{B.60}$$

Now, since $\ell_q(c) \to \infty$ as $c \to \infty$, and $\ell_q(\cdot)$ is continuous away from 0, there must be a point $c_{\max} > 0$ such that

$$\ell_q(c_{\max}) = \bar{v}, \text{ and} \tag{B.61}$$

$$\ell_q(c) \geq \bar{v} \quad \text{for all } c \geq c_{\max}. \tag{B.62}$$

This shows that $c^\star(\rho_q(\xi))$ cannot be greater than $c_{\max}$ when ξ is large, for otherwise (B.62) and (B.57) imply

$$\bar{v} \leq \ell(c^\star(\rho_q(\xi))) \tag{B.63}$$

$$\leq v(c^\star(\rho_q(\xi)), \rho_q(\xi)) \tag{B.64}$$

$$\leq v(c^\star(\bar{\rho}_q(\xi)), \rho_q(\xi)) \quad \text{by definition of } c^\star(\cdot), \tag{B.65}$$

contradicting line (B.60). Hence, line (B.56) is true.

Since we know that $c^\star(\rho_q(\xi))$ is a bounded sequence, we can show that $c^\star(\rho_q(\xi))$ converges to $c^\star(\bar{\rho}_q))$ if all of its convergent subsequences do. Likewise, suppose there is some $\breve{c} \in [\varepsilon_q, c_{\max}]$, such that along some subsequence $\xi_j \to \infty$,

$$c^\star(\rho_q(\xi_j)) \to \breve{c} \in [\varepsilon_q, c_{\max}]. \tag{B.66}$$

We now argue that $\breve{c}$ must be equal to $c^\star(\bar{\rho}_q)$. Due to the continuity of $\tilde{v}_q(\cdot,\cdot)$ and the limit (B.66), we have

$$\tilde{v}_q(\breve{c},\bar{\rho}_q) = \lim_{j\to\infty} \tilde{v}_q(c^\star(\rho_q(\xi_j)),\rho_q(\xi_j)) \tag{B.67}$$

$$\leq \lim_{j\to\infty} \tilde{v}_q(c^\star(\bar{\rho}_q),\rho_q(\xi_j)) \quad \text{by definition of } c^\star(\cdot) \tag{B.68}$$

$$= \tilde{v}_q(c^\star(\bar{\rho}_q),\bar{\rho}_q)) \tag{B.69}$$

$$\leq \tilde{v}_q(\breve{c},\bar{\rho}_q), \quad \text{by definition of } c^\star(\cdot). \tag{B.70}$$

Comparing the first line and the last line forces $\tilde{v}_q(\breve{c},\bar{\rho}_q) = \tilde{v}_q(c^\star(\bar{\rho}_q),\bar{\rho}_q)$, and so the uniqueness assumption **A3.4** gives $\breve{c} = c^\star(\bar{\rho}_q)$, as desired. □

B.5 Proofs for Section 3.6

Proof of Lemma 3.3.

Proof of inequality (3.72). It is enough to prove the result for $s_2(x)$ since $s_q(x) \geq s_2(x)$ for all $q \in [0,2]$. Let d be the dimension of the null space of A, and let $B \in \mathbb{R}^{p\times d}$ be a matrix whose columns are an orthonormal basis for the null space of A. If $x \neq 0$, then define the scaled matrix $\tilde{B} := \|x\|_\infty B$. (If $x = 0$, the steps of the proof can be repeated using $\tilde{B} = B$.) Letting $z \in \mathbb{R}^d$ be a standard Gaussian vector, we will study the random vector

$$\tilde{x} := x + \tilde{B}z,$$

which satisfies $Ax = A\tilde{x}$ for all realizations of z. We begin the argument by defining a function $f : \mathbb{R}^p \to \mathbb{R}$ according to

$$f(\tilde{x}) := \|\tilde{x}\|_1 - c(n,p)\|\tilde{x}\|_2, \tag{B.71}$$

where

$$c(n,p) := \tfrac{1}{\sqrt{\pi e}} \tfrac{(p-n)}{\sqrt{p}}. \tag{B.72}$$

The essential point to notice is that the event $\{f(\tilde{x}) > 0\}$ is equivalent to

$$\tfrac{\|\tilde{x}\|_1^2}{\|\tilde{x}\|_2^2} > c(n,p)^2 = \tfrac{1}{\pi e}(1-\tfrac{n}{p})^2 p,$$

which is the desired bound. (Note that $\tilde{x}$ is non-zero with probability 1.) Hence, a vector $\tilde{x}$ satisfying the bound (3.72) exists if the event $\{f(\tilde{x}) > 0\}$ occurs with positive probability. We will prove that the probability $\mathbb{P}(f(\tilde{x}) > 0)$ is positive by showing $\mathbb{E}[f(\tilde{x})] > 0$, which in turn can be reduced to upper-bounding $\mathbb{E}\|\tilde{x}\|_2$, and lower-bounding $\mathbb{E}\|\tilde{x}\|_1$. The upper bound on $\mathbb{E}\|\tilde{x}\|_2$ follows from Jensen's inequality and a direct calculation,

$$\begin{aligned}\mathbb{E}\|\tilde{x}\|_2 &= \mathbb{E}\|x + Bz\|_2\\ &< \sqrt{\mathbb{E}\|x + \tilde{B}z\|_2^2}\\ &= \sqrt{\|x\|_2^2 + \|\tilde{B}\|_F^2}\\ &= \sqrt{\|x\|_2^2 + \|x\|_\infty^2 d}.\end{aligned} \tag{B.73}$$

The lower bound on $\mathbb{E}\|\tilde{x}\|_1$ is more involved. If we let $\tilde{b}_i$ denote the ith row of $\tilde{B}$, then the ith coordinate of $\tilde{x}$ can be written as $\tilde{x}_i = x_i + \langle \tilde{b}_i, z\rangle$, which is distributed according to $N(x_i, \|\tilde{b}_i\|_2^2)$. Taking the absolute value $|\tilde{x}_i|$ results in a "folded normal" distribution, whose expectation can be calculated exactly as

$$\mathbb{E}|\tilde{x}_i| = \|\tilde{b}_i\|_2 \sqrt{\tfrac{2}{\pi}} \exp\left(\tfrac{-x_i^2}{2\|\tilde{b}_i\|_2^2}\right) + |x_i|\Big(1 - 2\Phi\big(\tfrac{-|x_i|}{\|\tilde{b}_i\|_2}\big)\Big), \tag{B.74}$$

where Φ is the standard normal distribution function. Note that it is possible to have $\|\tilde{b}_i\|_2 = 0$, in which case $\tilde{x}_i = x_i$. This separate case can be easily handled in the rest of the argument.

When $|x_i|/\|\tilde{b}_i\|_2$ is small, the first term on the right side of (B.74) dominates, and then $\mathbb{E}|\tilde{x}_i|$ is roughly $\|\tilde{b}_i\|_2$. Alternatively, when $|x_i|/\|\tilde{b}_i\|_2$ is large, the second term dominates, and then $\mathbb{E}|\tilde{x}_i|$ is roughly $|x_i|$. Thus, it is natural to consider the set of indices $\mathcal{I}_1 = \{i : \|\tilde{b}_i\|_2 \geq |x_i|\}$, and its complement $\mathcal{I}_2 = \{i : \|\tilde{b}_i\|_2 < |x_i|\}$. This leads us to the following bounds,

$$\begin{aligned}\mathbb{E}\|\tilde{x}\|_1 &= \sum_{i\in\mathcal{I}_1} \mathbb{E}|\tilde{x}_i| + \sum_{i\in\mathcal{I}_2} \mathbb{E}|\tilde{x}_i|\\ &\geq \sum_{i\in\mathcal{I}_1} \|\tilde{b}_i\|_2 \sqrt{\tfrac{2}{\pi}} \exp(-\tfrac{1}{2}) + \sum_{i\in\mathcal{I}_2} |x_i|(1 - 2\Phi(-1))\\ &\geq \sum_{i\in\mathcal{I}_1} \|\tilde{b}_i\|_2 \sqrt{\tfrac{2}{\pi e}} + \sum_{i\in\mathcal{I}_2} \|\tilde{b}_i\|_2(1 - 2\Phi(-1))\\ &\geq \sqrt{\tfrac{2}{\pi e}} \sum_{i=1}^p \|\tilde{b}_i\|_2 \qquad \text{using } (1 - 2\Phi(-1)) \geq \sqrt{\tfrac{2}{\pi e}},\\ &= \sqrt{\tfrac{2}{\pi e}} \|x\|_\infty \sum_{i=1}^p \|b_i\|_2,\end{aligned} \tag{B.75}$$

where b_i is the ith row of $B \in \mathbb{R}^{p\times d}$. Since the matrix $B \in \mathbb{R}^{p\times d}$ has orthonormal columns, it may be regarded as a submatrix of an orthogonal $p \times p$ matrix, and so the rows b_i satisfy $\|b_i\|_2 \leq 1$, yielding $\|b_i\|_2 \geq \|b_i\|_2^2$. Hence,

$$\textstyle\sum_{i=1}^p \|b_i\|_2 \geq \sum_{i=1}^p \|b_i\|_2^2 = \|B\|_F^2 = d \geq p - n.$$

Altogether, we obtain the bound

$$\mathbb{E}\|\tilde{x}\|_1 \geq \sqrt{\tfrac{2}{\pi e}}\|x\|_\infty(p-n). \tag{B.76}$$

Combining the bounds (B.76) and (B.73), and noting that $d \leq p$, we obtain

$$\begin{aligned} \frac{\mathbb{E}\|\tilde{x}\|_1}{\mathbb{E}\|\tilde{x}\|_2} &> \frac{\sqrt{\frac{2}{\pi e}}\|x\|_\infty(p-n)}{\sqrt{\|x\|_2^2+\|x\|_\infty^2 p}} \\ &= \frac{\sqrt{\frac{2}{\pi e}}(p-n)}{\sqrt{\frac{\|x\|_2^2}{\|x\|_\infty^2}+p}} \\ &\geq \tfrac{1}{\sqrt{\pi e}}\tfrac{(p-n)}{\sqrt{p}} \\ &= c(n,p), \end{aligned} \tag{B.77}$$

where we have used the fact that $\frac{\|x\|_2^2}{\|x\|_\infty^2} \leq p$. This proves $\mathbb{E}[f(\tilde{x})] > 0$, giving (3.72). □

Proof of inequality (3.73). It is enough to prove the result for $s_\infty(x)$ since $s_q(x) \geq s_\infty(x)$ for all $q \in [0,\infty]$. We retain the same notation as in the proof above. Following the same general argument, it is enough to show that

$$\frac{\mathbb{E}\|\bar{x}\|_1}{\mathbb{E}\|\bar{x}\|_\infty} > \bar{c}(n,p), \tag{B.78}$$

where

$$\bar{c}(n,p) := \frac{\sqrt{\frac{2}{\pi e}}(p-n)}{1+\sqrt{16\log(2p)}}. \tag{B.79}$$

In particular, we will re-use the bound

$$\mathbb{E}\|\tilde{x}\|_1 \geq \sqrt{\tfrac{2}{\pi e}}\|x\|_\infty(p-n). \tag{B.80}$$

The new item to handle is an upper bound on $\mathbb{E}\|\tilde{x}\|_\infty$. Clearly, we have $\|\tilde{x}\|_\infty \leq \|x\|_\infty + \|\tilde{B}z\|_\infty$, and so it is enough to upper-bound $\mathbb{E}\|\tilde{B}z\|_\infty$. We will do this using a version of Slepian's inequality. If $\tilde{b}_i$ denotes the i^{th} row of $\tilde{B}$, define the random variable $g_i = \langle \tilde{b}_i, z\rangle$, and let $w_1,\ldots,w_p$ be i.i.d. $N(0,1)$ variables. The idea is to compare the Gaussian process g_i with the Gaussian process $\sqrt{2}\|x\|_\infty w_i$. By Proposition A.2.6 in the book [VW96], the inequality

$$\mathbb{E}\|\tilde{B}z\|_\infty = \mathbb{E}\left[\max_{1\leq i\leq p}|g_i|\right] \leq 2\sqrt{2}\|x\|_\infty\,\mathbb{E}\left[\max_{1\leq i\leq p}|w_i|\right],$$

holds as long as the condition $\mathbb{E}(g_i - g_j)^2 \leq 2\|x\|_\infty^2 \, \mathbb{E}(w_i - w_j)^2$ is satisfied for all $i, j \in \{1, \dots, p\}$. This can be verified by first noting that $g_i - g_j = \langle \tilde{b}_i - \tilde{b}_j, z \rangle$, which is distributed according to $N(0, \|\tilde{b}_i - \tilde{b}_j\|_2^2)$. Since $\|\tilde{b}_i\|_2 \leq \|x\|_\infty$ for all i, it follows that

$$\begin{aligned} \mathbb{E}(g_i - g_j)^2 &= \|\tilde{b}_i - \tilde{b}_j\|_2^2 \\ &\leq 4\|x\|_\infty^2 \\ &= 2\|x\|_\infty^2 \mathbb{E}(w_i - w_j)^2, \end{aligned} \tag{B.81}$$

as needed. To finish the proof, we make use of a standard bound for the expectation of Gaussian maxima

$$\mathbb{E}\left[\max_{1 \leq i \leq p} |w_i|\right] < \sqrt{2\log(2p)},$$

which follows from a modification of the proof of Massart's finite class lemma [Mas00, Lemma 5.2]. Combining the last two steps, we obtain

$$\mathbb{E}\|\tilde{x}\|_\infty < \|x\|_\infty + 2\sqrt{2}\|x\|_\infty \sqrt{2\log(2p)}. \tag{B.82}$$

Hence, the bounds (B.80) and (B.82) clearly lead to (B.78). □

Proof of Theorem 3.4

Proof. We begin by making some reductions. First, we claim it is enough to show that

$$\inf_{A \in \mathbb{R}^{n \times p}} \inf_{\delta : \mathbb{R}^n \to \mathbb{R}} \sup_{x \in \mathbb{R}^p \setminus \{0\}} \Big| \delta(Ax) - s_2(x) \Big| \geq \tfrac{1}{2} \tfrac{1}{\pi e} \cdot (1 - \tfrac{n}{p})^2 \cdot p - \tfrac{1}{2}. \tag{B.83}$$

To see this, note that the general inequality $s_2(x) \leq p$ implies

$$\left| \tfrac{\delta(Ax)}{s_2(x)} - 1 \right| \geq \tfrac{1}{p} \left| \delta(Ax) - s_2(x) \right|,$$

and we can optimize both sides with respect x, δ, and A. Next, for any fixed matrix $A \in \mathbb{R}^{n \times p}$, it is enough to show that

$$\inf_{\delta : \mathbb{R}^n \to \mathbb{R}} \sup_{x \in \mathbb{R}^p \setminus \{0\}} \Big| \delta(Ax) - s_2(x) \Big| \geq \tfrac{1}{2} \tfrac{1}{\pi e} \cdot (1 - \tfrac{n}{p})^2 \cdot p - \tfrac{1}{2}, \tag{B.84}$$

as we may take the infimum over all matrices A without affecting the right hand side. To make a third reduction, it is enough to prove the same bound when $\mathbb{R}^p \setminus \{0\}$ is replaced with any subset, as this can only make the supremum smaller. In particular, we replace $\mathbb{R}^p \setminus \{0\}$ with the two-point subset $\{e_1, \tilde{x}\}$, where $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^p$, and by Lemma 1, there exists $\tilde{x}$ to satisfying $Ae_1 = A\tilde{x}$, with

$$s_2(e_1) = 1, \quad \text{and} \quad s_2(\tilde{x}) \geq \tfrac{1}{\pi e} \cdot (1 - \tfrac{n}{p})^2 \cdot p.$$

We now complete the proof by showing that the lower bound (B.84) holds for the two-point problem, i.e.

$$\inf_{\delta:\mathbb{R}^n\to\mathbb{R}} \sup_{x\in\{e_1,\tilde{x}\}} \Big|\delta(Ax) - s_2(x)\Big| \geq \tfrac{1}{2}\tfrac{1}{\pi e}\cdot(1-\tfrac{n}{p})^2\cdot p - \tfrac{1}{2}, \tag{B.85}$$

and we will accomplish this using the classical technique of constructing a Bayes procedure with constant risk. For any decision rule $\delta : \mathbb{R}^n \to \mathbb{R}$, any $A \in \mathbb{R}^{n\times p}$, and any point $x \in \{e_1, \tilde{x}\}$, define the (deterministic) risk function

$$R(x,\delta) := \big|\delta(Ax) - s_2(x)\big|.$$

Also, for any prior π on the two-point set $\{e_1, \tilde{x}\}$, define

$$r(\pi,\delta) := \int R(x,\delta)d\pi(x).$$

By Propositions 3.3.1 and 3.3.2 of [BD01], the inequality (B.85) holds if there exists a prior distribution π^* on $\{e_1,\tilde{x}\}$ and a decision rule $\delta^* : \mathbb{R}^n \to \mathbb{R}$ with the following three properties:

1. The rule δ^* is Bayes for π^*, i.e. $r(\pi^*,\delta^*) = \inf_\delta r(\pi^*,\delta)$.
2. The rule δ^* has constant risk over $\{e_1,\tilde{x}\}$, i.e. $R(e_1,\delta^*) = R(\tilde{x},\delta^*)$.
3. The constant value of the risk of δ^* is at least $\frac{1}{2}\frac{1}{\pi e}\cdot(1-\frac{n}{p})^2\cdot p - \frac{1}{2}$.

To exhibit π^* and δ^* with these properties, we define π^* to be the two-point prior that puts equal mass at e_1 and $\tilde{x}$, and we define δ^* to be the trivial decision rule that always returns the average of the two possibilities, namely $\delta^*(Ax) \equiv \frac{1}{2}(s_2(\tilde{x}) + s_2(e_1))$ for all $x \in \{e_1,\tilde{x}\}$. It is simple to check the second and third properties. To check that δ^* is Bayes for π^*, the triangle inequality gives

$$\begin{aligned} r(\pi^*,\delta) &= \tfrac{1}{2}\Big|\delta(A\tilde{x}) - s_2(\tilde{x})\Big| + \tfrac{1}{2}\Big|\delta(Ae_1) - s_2(e_1)\Big|, \\ &\geq \tfrac{1}{2}\big|s_2(\tilde{x}) - s_2(e_1)\big| \\ &= \tfrac{1}{2}\big|\delta^*(A\tilde{x}) - s_2(\tilde{x})\big| + \tfrac{1}{2}\big|\delta^*(Ae_1) - s_2(e_1)\big| \\ &= r(\pi^*,\delta^*), \end{aligned} \tag{B.86}$$

which holds for every δ, implying that δ^* is Bayes for π^*. □

B.6 Background results

Convergence of medians for triangular array

Suppose $F_n \to_w F$. Let $\mathbb{F}_n$ be the empirical distribution corresponding to F_n. With regard to the proof of Proposition 3.3, we would like to show that $|\text{med}(\mathbb{F}_n) - \text{med}(F_n)| \to 0$ when

F has a continuous cdf. We will take our median to be $\mathbb{F}_n^{-1}(1/2)$, i.e. the empirical quantile process evaluated at $1/2$. Also note that the distribution function F arising in the proof of Proposition 3.3 is continuous because it arises from convolution with a stable law.

Lemma B.2. *Suppose F is continuous and $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$. Then*

$$|\mathbb{F}_n^{-1}(\tfrac{1}{2}) - F_n^{-1}(\tfrac{1}{2})| \to 0, \quad a.s.$$

Proof. Let $\mathbb{G}_n$ be the empirical c.d.f. corresponding to a sample of n i.i.d. uniform variables on $[0,1]$. First note that

$$|\mathbb{F}_n^{-1}(\tfrac{1}{2}) - F_n^{-1}(\tfrac{1}{2})| = |F_n^{-1}(\mathbb{G}_n^{-1}(\tfrac{1}{2})) - F_n^{-1}(\tfrac{1}{2})|$$

It is a basic fact that $\mathbb{G}_n^{-1}(\frac{1}{2}) \to \frac{1}{2}$ almost surely. Hence, it suffices to show that $|F_n^{-1}(a_n) - F_n^{-1}(a)| \to 0$ for any real sequence $a_n \to a$. To see this, we write

$$|F_n^{-1}(a_n) - F_n^{-1}(a)| \leq |F_n^{-1}(a_n) - F^{-1}(a_n)| + |F^{-1}(a_n) - F^{-1}(a)| + |F^{-1}(a) - F_n^{-1}(a)|.$$

Since F_n^{-1} is a sequence of monotone functions converging to a continuous function F^{-1}, it is a basic fact that F_n^{-1} converges uniformly to F^{-1} on compact subsets. Hence, the first term on the right hand side tends to 0. It is obvious that the other terms tend to 0.

A unique minimizer for the variance function with stable noise

In this subsection, we aim to show that when the noise distribution is $\text{stable}_q(1)$, the variance function $\tilde{v}_q(\cdot, \bar{\rho}_q)$ has a unique minimizer in $[\varepsilon_q, \infty)$.[5] Note that since φ_0 has no roots in this case, the extended variance function $\tilde{v}_q(c, \bar{\rho}_q)$ agrees with $v_q(c, \bar{\rho}_q)$ for all $c \neq 0$. Furthermore, when $q \in (0,2)$ the minimizer cannot occur at $c = 0$ due to Lemma 3.2. Hence, when $q \in (0,2)$ it is enough to check that $v_q(\cdot, \bar{\rho}_q)$ has a unique minimizer in $(0, \infty)$.

Recall that the characteristic function for $\text{stable}_q(1)$ is

$$\varphi_0(t) = \exp(-|t|^q), \tag{B.87}$$

and it follows that for any $q \in (0, 2]$, the variance function is given by

$$\begin{aligned} v_q(c, \bar{\rho}_q) &= \tfrac{1}{|c|^{2q}} \Big(\tfrac{1}{2} \tfrac{1}{\varphi_0(\bar{\rho}_q |c|)^2} \exp(2|c|^q) + \tfrac{1}{2} \tfrac{\varphi_0(2\bar{\rho}_q|c|)}{\varphi_0(\bar{\rho}_q|c|)^2} \exp((2 - 2^q)|c_0|^q) - 1 \Big) \\ &= \tfrac{1}{|c|^{2q}} \Big(\tfrac{1}{2} \exp((2(\bar{\rho}_q^q + 1)|c|^q) + \tfrac{1}{2} \exp((2 - 2^q)(\bar{\rho}_q^q + 1)|c|^q) - 1 \Big). \end{aligned} \tag{B.88}$$

Now consider the monotone change of variable $u := |c|^q$, and notice that $v_q(c, \bar{\rho}_q) = f(u)/u^2$ where

$$f(u) := \tfrac{1}{2} \exp(2(\bar{\rho}_q^q + 1)u) + \tfrac{1}{2} \exp((2 - 2^q)(\bar{\rho}_q^q + 1)u) - 1. \tag{B.89}$$

The following lemma demonstrates the desired claim by showing that $u \mapsto f(u)/u^2$ is strictly convex on $(0, \infty)$. (We omit the simple derivative calculations involved in checking that $f(u)$ satisfies the conditions of the lemma.)

[5] Recall that $\varepsilon_q = 0$ for $q \in (0,2)$ and $\varepsilon_2 > 0$.

Lemma B.3. *Let $f : [0, \infty) \to \mathbb{R}$ be a 4-times differentiable function such that $f(0) \geq 0$, $f'(0) \geq 0$, and $f^{(4)}(u) > 0$ for all $u > 0$. Then, the function $u \mapsto \frac{f(u)}{u^2}$ is strictly convex on $(0, \infty)$.*

Proof. Let $h(u) = \frac{f(u)}{u^2}$, and let $\psi(u) = u^4 h''(u)$. To show that $h''(u)$ is strictly positive on $(0, \infty)$, it suffices to show that $\psi(u) > 0$ for all $u > 0$. By direct calculation,

$$\psi(u) = u^2 f''(u) - 4uf'(u) + 6f(u),$$

and so the assumption $f(0) \geq 0$ implies $\psi(0) \geq 0$. Consequently, it is enough to show that ψ is strictly increasing on $(0, \infty)$. Since,

$$\psi'(u) = u^2 f^{(3)}(u) - 2uf''(u) + 2f'(u),$$

the assumption $f'(0) \geq 0$ implies $\psi'(0) \geq 0$, and so it is enough to show that ψ' is strictly increasing on $(0, \infty)$. Differentiating ψ' leads to a notable cancellation, giving

$$\psi''(u) = u^2 f^{(4)}(u),$$

and so the assumption on $f^{(4)}(u) > 0$ for all $u > 0$ completes the proof. □