High-Dimensional Statistics: A Non-Asymptotic Viewpoint

Chapter 7: Sparse linear models in high dimensions

Notes by Renxiong Liu

In this chapter, we discuss the most basic problem in high-dimensional statistics, which is the so-called high dimensional linear regression. In its low-dimensional setting, in which the number of predictors d is substantially less than the sample size n, the associated theory is classical. By contrast, the goal in this chapter is to develop theory applicable to the high-dimensional regime: (1) $n \asymp p$ and (2) $p \gg n$. As one might expect, if the model lacks any additional structure, then there is no hope of obtaining consistent estimator when p/n not goes to 0. This leads to the great interest in sparsity, we focuses on different types of sparse models in this chapter.

7.1 Problem Settings

Let $\theta^\star \in \mathbb{R}^d$ be the "true" parameter, $X \in \mathbb{R}^{n\times d}$ be the feature matrix and $Y \in \mathbb{R}^n$ be the response vector. The model we consider is in following form:

$$Y = X\theta^\star + \epsilon \tag{7.1}$$

The focus of this chapter is settings in which the sample size n is smaller than the number of predictors d. Note when $n < d$, it is impossible to obtain any meaningful estimates of $\theta^\star$ unless there is some additional assumption on the low-dimensional structure. One widely considered assumption is the *hard sparsity* assumption:

Assumption 7.1 (hard sparsity)**.** *Denote by $S(\theta^\star) = \{j \in \{1, 2 \ldots, d\} \mid \theta^\star_j \neq 0\}$ the support of vector $\theta^\star$. We assume that the cardinality $|S(\theta^\star)|$ is substantially smaller than d, i.e. $|S(\theta^\star)| \leq d' \ll d$.*

Note Assumption 7.1 is a bit strong since it regularizes the maximal number of non-zero terms to be some value substantially smaller than d. In real applications, we might prefer to use another notion of sparsity, named weak sparsity, where lots of terms are closed to zero. There are different ways to formalize such an idea, one is via l_q-norm, which is formalized in the following assumption.

Assumption 7.2 (weak sparsity)**.** *$\theta^\star$ lies in the l_q-ball around 0, that is,*

$$B_q(r_q) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_q \leq r_q\} \tag{7.2}$$

Examples of Assumption 7.2 are given in Figure 7.1.

7.2 Applications of sparse linear models

Before studying the theoretical performance of estimators under sparse assumption, we first introduce some applications of sparse linear model.

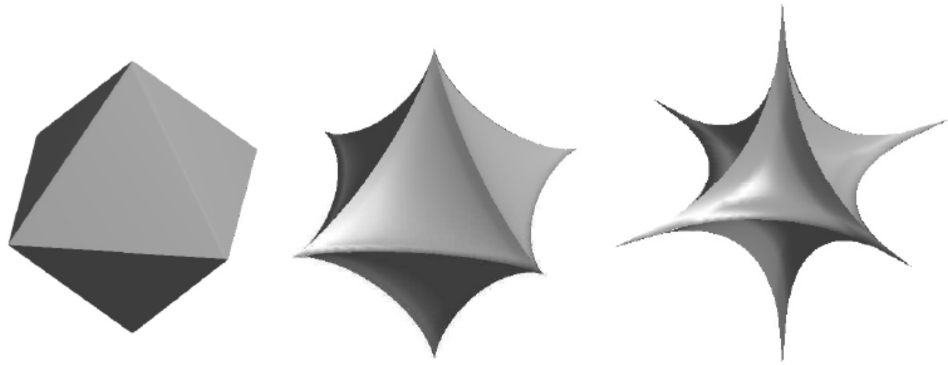

Figure 7.1: Illustrations of the l_q balls for different choices of parameter $q \in (0, 1]$. (a) $q = 1$, (b) $q = 0.75$ and (c) $q = 0.5$.

Example 7.1 (Gaussian sequence model)**.** We assume the observations are generated from the following model sequentially:

$$Y_k = \sqrt{n}\theta_k^\star + \epsilon_k \tag{7.3}$$

where $\epsilon_k \sim N(0, \sigma^2)$ are i.i.d. and $k = 1, \ldots, n$. Note this model can be rewritten as

$$Y = \sqrt{n}I\theta^\star + \epsilon \tag{7.4}$$

One important characteristic of the model above is, the number of parameters $p = n$, which means the number of parameters grows as the number of observations grows. Althoughit appears simple on the surface, it is a surprisingly rich model: indeed, many problems in nonparametric estimation, among them regression and density estimation, can be reduced to an "equivalent" instance of the Gaussian sequence model, in the sense that the optimal rates for estimation are the same under both models.

Example 7.2 (Selection of Gaussian graphical models)**.** Let $(X_1, X_2, \ldots, X_n)$ be a zero-mean Gaussian random vector with a non-degenerate covariance matrix. The density function is represented as

$$p(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^d \det(\Theta^{-1})}} \exp(-\frac{1}{2}x^\top \Theta x) \tag{7.5}$$

where Θ is the precision-matrix. By theory of Gaussian graphical model, we have that (i, j)-term in Θ is zero if and only if $x_i \perp x_j \mid x_{-\{i,j\}}$, which means there is no linkage between node i and j in graph representation. Hence it's important to introduce some assumptions on the sparsity of precision matrix Θ.

7.3 Recovery in the noiseless setting

In this section, we consider the model that observations are perfect, i.e., we consider

$$Y = X\theta \tag{7.6}$$

Assume we are told that there is some vector $\theta^\star$ with at most $s \ll d$ non-zero entries such that $y = X\theta^\star$. Our goal is to recover the sparse solution to the linear system. That is, we aim to solve problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0, \text{ such that } X\theta = Y \tag{7.7}$$

where $\|\theta\|_0 = \sum_{k=1}^d \mathbb{I}(\theta_k \neq 0)$ is the psedo-norm that represents the number of non-zero entries of vector θ. Note the object function is non-smooth and non-convex, which is computationally intractable. The most closed convex relaxation is to replace the l_0 norm by l_1 norm, that is,

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \text{ such that } X\theta = Y \tag{7.8}$$

We refer this problem as basis pursuit linear program (see Chen, Donoho and Saunders (1998))

7.3.1 Exact recovery and restricted nullspace

We now turn to a theoretical question: when is solving the basis pursuit program (l_1 problem) (7.8) equivalent to solving the original l_0-problem (7.7).

Assume that $\theta^\star \in \mathbb{R}^d$ such that $Y = X\theta^\star$, which means the linear restriction is feasible. Next, we denote by $S = \{j \in \{1, 2 \ldots, d\} \mid \theta_j^\star \neq 0\}$ the support vector and S^c the complement of S. Recall any solution to $X\theta = Y$ must be of form

$$\theta^\star + \Delta, \text{ where } \Delta \in \textbf{Null}(X) = \{\Delta \mid X\Delta = 0\} \tag{7.9}$$

To proceed, we define the tangent cone of the l_1-ball at $\theta^\star$, that is,

$$T(\theta^\star) = \{\Delta \in \mathbb{R}^d \mid \|\theta^\star + t\Delta\|_1 \leq \|\theta^\star\|_1, \text{ for some } t > 0\} \tag{7.10}$$

It's easy to verify that $T(\theta^\star)$ is a cone. For a tangent cone, the most important information is the direction. Note $T(\theta^\star) \cap \textbf{Null}(X)$ is still a cone, meaning that if $T(\theta^\star) \cap \textbf{Null}(X) \neq \emptyset$, then $\theta^\star$ is not a solution to basis pursuit problem (7.8). Here we present two possible cases between the solution of (7.8) and $\theta^\star$ (see Figure 7.2).

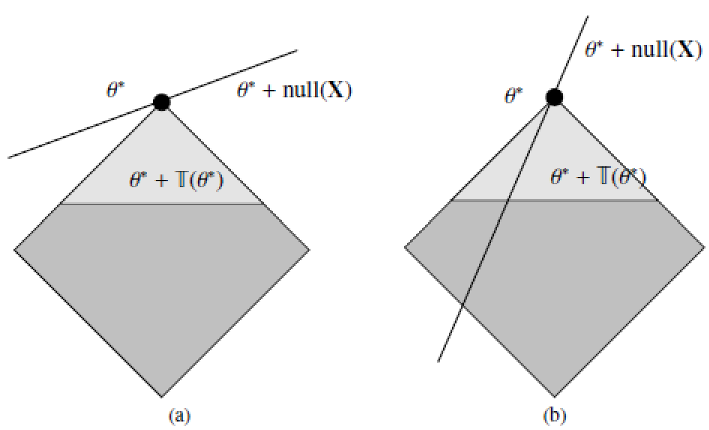

Figure 7.2: Two possible cases for whether $\theta^\star$ is the unique solution.

Next, we investigate the condition for $\theta^\star$ to be the unique solution of problem (7.8), which is known as the *restricted nullspace property*. More specifically, given support S, we define

$$C(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\} \tag{7.11}$$

We have the following definition for restricted nullspace property.

Definition 7.1 (restricted nullspace property)**.** Matrix X satisfies the restricted nullspace property with respect to S, if $C(S) \cap \mathbf{Null}(X) = \{0\}$.

One quick way to understand restricted nullspace property is: let $\theta^\star = (\theta_1, \dots, \theta_s, 0, \dots, 0)$ be the underlying true vector. In order to make $\theta^\star$ possible to be recovered exactly, the optimization problem (7.8) should have unique solution $\theta^\star$. This means for any nonzero $\Delta = (\Delta_1, \dots, \Delta_d) \in \mathbf{Null}(X)$, $\theta^\star + \Delta = (\theta_1 + \Delta_1, \dots, \theta_s + \Delta_s, \Delta_{s+1}, \dots, \Delta_d)$ cannot have smaller l_1 norm than $\theta^\star = (\theta_1, \dots, \theta_s, 0, \dots, 0)$. Hence

$$\sum_{i \in S^c} |\Delta_i| > \sum_{i \in S} |\theta_i| - \sum_{i \in S} |\theta_i + \Delta_i| \tag{7.12}$$

Since $X\Delta = X(-\Delta) = 0$, we have that

$$\sum_{i \in S^c} |\Delta_i| \geq \max\{\sum_{i \in S}(|\theta_i| - |\theta_i + \Delta_i|), \sum_{i \in S}(|\theta_i| - |\theta_i - \Delta_i|)\} \tag{7.13}$$

For a small enough Δ, (i.e., $|\Delta_i| < |\theta_i|$ for $i \in S$), this implies $\|\Delta_{S^c}\|_1 > \|\Delta_S\|_1$. Hence there is no interception between $C(S)$ and $\mathbf{Null}(X)$.

The following theorem justifies above arguments in more details.

Theorem 7.1. *The following two properties are equivalent:*

a. For any vector $\theta^\star$ with support S, the basis pursuit program (7.8) applied with condition $X\theta^\star = Y$ has unique solution $\hat{\theta} = \theta^\star$.

b. The matrix X satisfies the restricted nullspace property with respect to S.

Proof. We first show that $b \Rightarrow a$. Denote by $\hat{\theta}$ a solution of problem (7.8). By optimality of $\hat{\theta}$, we have that $\|\hat{\theta}\|_1 \leq \|\theta^\star\|_1$. Let $\Delta = \hat{\theta} - \theta^\star$, note $X\Delta = 0$, by restricted nullspace property, we have either $\Delta = 0$ or $\|\Delta_{S^c}\|_1 > \|\Delta_S\|_1$ when $\Delta \neq 0$. For the second case, we have that $\|\theta^\star_S\|_1 = \|\theta^\star\|_1 \geq \|\hat{\theta}\|_1 \geq \|\Delta_{S^c}\|_1 + \|\theta^\star_S\|_1 - \|\Delta_S\|_1$, which implies $\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1$. Hence we get the contradiction. This measn $\Delta = 0$, or equivalently, $\theta^\star$ is the unique solution to the basis pursuit program (7.8), when condition $X\theta^\star = Y$ is equipped.

Next, we show $a \Rightarrow b$. By condition (a.), for any $\Delta \in \mathbf{Null}(X)$ with $\Delta \neq 0$, $(\Delta_S, 0)^\top$ is the unique solution of the following problem:

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \text{ such that } X\theta = X(\Delta_S, 0)^\top \tag{7.14}$$

Note $X(0, -\Delta_{S^c})^\top - X(\Delta_S, 0)^\top = -X\Delta = 0$, we must have that $(0, -\Delta_{S^c})^\top$ is a feasible solution. By optimality of $(\Delta_S, 0)^\top$, this means $\|\Delta_S\|_1 < \|\Delta_{S^c}\|_1$. Hence $\Delta \notin C(S)$, hich means X satisfies the restricted nullspace property with respect to S.

This completes the proof.

Now we have know that, exact recovery is equivalent to the requirement that X satisfies the restricted nullspace property with respect to S. One question is: when will X satisfies the restricted nullspace property with respect to S? Next, we give some sufficient conditions for restricted nullspace property to hold.

The earliest sufficient condition is on the pairwise *incoherence parameter*, which is defined as

$$\delta_{pw}(X) = \max_{i,j} |(\frac{1}{n}X^\top X - I)_{i,j}| \tag{7.15}$$

The smaller $\delta_{pw}(X)$ is, the more $\frac{1}{\sqrt{n}}X$ is closed to an orthonormal matrix. The next theorem guarantee a uniform version of the restricted nullspace property.

Theorem 7.2. *If the pairwise incoherence satisfies the bound $\delta_{pw}(X) \leq \frac{1}{3s}$, then the restricted nullspace property holds for all subsets S of cardinality at most s.*

The proof will be uploaded in later version.

A more widely used sufficient condition is by the restricted isometry property (RIP), which is a natural generalization of the pairwise incoherence condition.

Definition 7.2 (Restricted isometry property)**.** For a given integer $s \in \{1, 2, \ldots, d\}$, we say that $X \in \mathbb{R}^{n\times d}$ satisfies a restricted isometry property of order s with constant $\delta_s(X) > 0$, if

$$\|\frac{1}{n}X_s^\top X_s - I\|_2 \leq \delta_s(X) \tag{7.16}$$

for all subsets S of size at most s. Here $\|A\|_2 = \sigma_{\max}(A)$ is the largest singular value.

Definition 7.2 implies that, for any subsets S of size at most s, $\frac{1}{n}X_s^\top X_s - I \preceq \delta_s^2(X)I$, or equivalently

$$\frac{1}{n}X^\top X - I \preceq \delta_s^2(X)I \tag{7.17}$$