Sparse linear models in high dimensions

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

Motivation

Classical vs High-Dimensional asymptotics

- Classical: low-dimensional settings, in which the number of predictors d is substantially less than the sample size n - i.e., $d \ll n$.
- High-dimensional: High-dimensional regime allows for scaling such that $d \asymp n$ or even $d \gg n$.

In the case that $d \gg n$, if the model lacks any additional structure, then there is no hope of obtaining consistent estimators when the ratio d/n stays bounded away from zero. Therefore, when working in settings in which $d > n$, it is necessary to impose additional structure on the unknown regression vector $\theta^* \in \mathbb{R}^d$.

Let $\theta^* \in \mathbb{R}^d$ be an unknown vector, and suppose we observe a vector $y \in \mathbb{R}^n$ and a matrix $X \in \mathbb{R}^{n \times d}$, such that $X = [x_1', \cdots, x_n']'$ that are linked via the linear model

$$y = X\theta^* + \varepsilon$$

where $\varepsilon \in \mathbb{R}^n$ is the noise vector. This model can be written in any of the following scalar forms

$$y_i = \langle x_i, \theta^* \rangle + \varepsilon_i, \quad i = 1, \cdots, n,$$
$$y_i = x_i'\theta + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $\langle x_i, \theta^* \rangle = \sum_{i=1}^{n} x_{ij}\theta_j^*$ denotes the Euclidean inner product.

The focus of this presentation is to consider the cases where $n < d$. We first consider the noiseless linear model, such that $\epsilon = 0$, in which we may model the response variable as

$$y = X\theta^*$$

which when $n < d$ defines an undetermined linear system, and the goal is to understand the structure of its sparse solutions.

When $d > n$, it is impossible to obtain any meaningful estimate of θ^* unless the model is equipped with some form of low-dimensional structure. First, we consider the simplest case, namely the hard sparsity assumption:

Hard sparsity assumption

The simplest kind of structure is the hard sparsity assumption that the set

$$S(\theta^*) := \{j \in \{1, \cdots, d\} \mid \theta_j^* \neq 0\}.$$

which is known as the support of θ^* and has cardinality $s := |S(\theta^*)|$, where $s \ll d.$,

The problem with the hard sparsity assumption is that it is overly restrictive, which motivates considering the weak sparsity assumption.

Definition

A vector θ^* is weakly sparse if it can be closely approximated by a sparse vector.

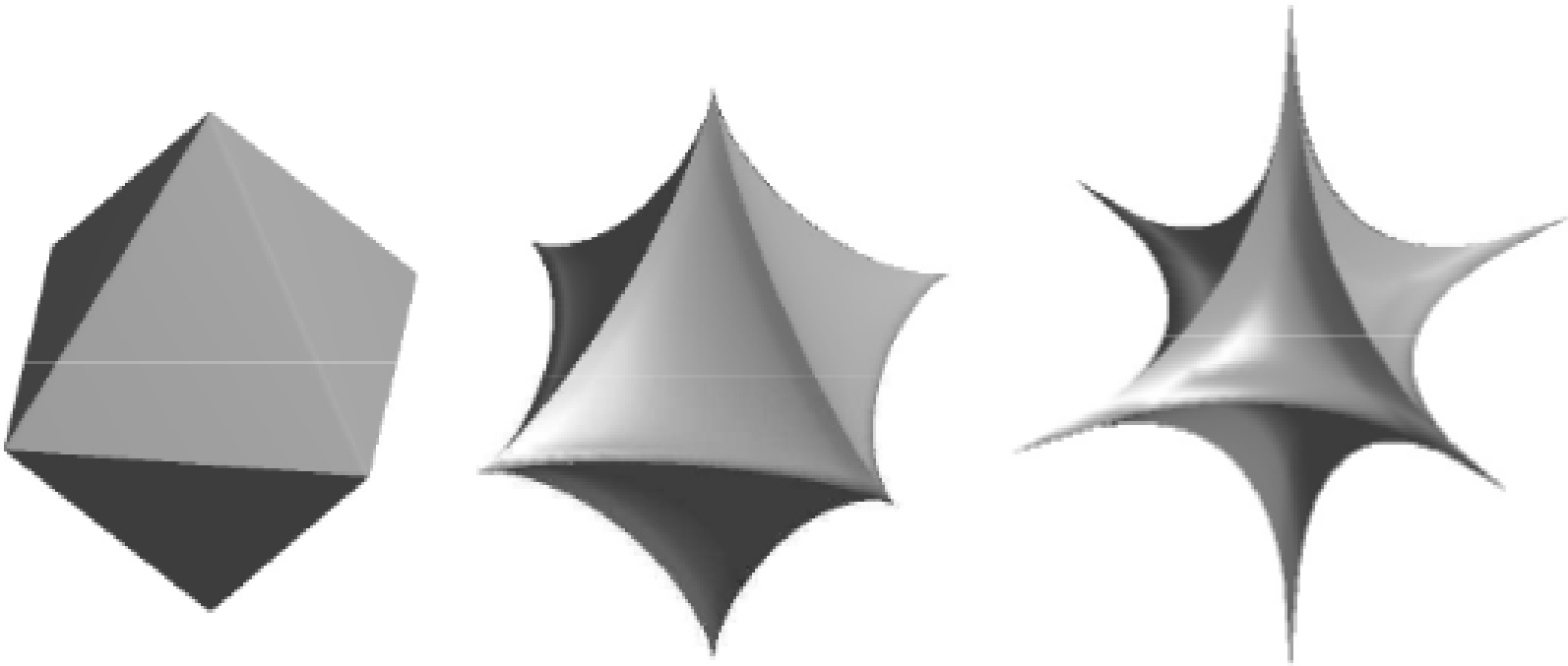
One way to formalize such an idea is via the l_q-norms. For a parameter $q \in [0, 1]$ and radius $R_1 > 0$, consider the l_q-ball set

$$B_q(R_q) = \left\{ \theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q \right\}$$

is one with radius R_q. As it is evident from the below figures for $q \in [0, 1)$, it is not a ball in the strict sense, since it a non-convex set. When $q = 0$, this is the case of the "improper" l_0-norm, and any vector $\theta^* \in B_0(R_0)$ can have at most $s = R_0$ non-zero entries. For values of $q \in (0, 1]$, membership in the set $B_q(R_q)$ has different interpretations, one of which involves, how quickly the ordered coefficients

$$|\theta^*_{(1)}| \geq |\theta^*_{(2)}| \geq \cdots \geq |\theta^*_{(d)}|$$

decay.



(a) (b) (c)

Figure 7.1 Illustrations of the ℓ_q-"balls" for different choices of the parameter $q \in (0, 1]$. (a) For $q = 1$, the set $\mathbb{B}_1(R_q)$ corresponds to the usual ℓ_1-ball shown here. (b) For $q = 0.75$, the ball is a non-convex set obtained by collapsing the faces of the ℓ_1-ball towards the origin. (c) For $q = 0.5$, the set becomes more "spiky", and it collapses into the hard sparsity constraint as $q \to 0^+$. As shown in Exercise 7.2(a), for all $q \in (0, 1]$, the set $\mathbb{B}_q(1)$ is star-shaped around the origin.

Example (Gaussian sequence models): Suppose we observed $\{y_1, \cdots, y_n\}$ where

$$y_i = \theta_i^* + \epsilon\varepsilon_i,$$

where $\varepsilon_i \sim N(0,1)$ and $\epsilon = \frac{\sigma}{\sqrt{n}}$, where the variance is divided by n, as it corresponds to taking n i.i.d variables and taking their average. In this case, it is evident that $n = d$ and as $n \to \infty$, so does $d \to \infty$. It is clearly evident that in the general linear model introduced earlier - i.e.

$$y = X\theta^* + \varepsilon$$

$X = I_n$.

Example (Signal denoising in orthonormal bases): Sparsity plays an important role in signal processing, both for compression and for denoising of signals. Suppose we have the noisy observations $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_n)'$.

$$\tilde{y} = \beta^* + \tilde{\varepsilon}$$

where the vector $\beta^* \in \mathbb{R}^d$ represents the signal, while $\tilde{\varepsilon}$ is some kind of additive noise. Denoising $\tilde{y}$ implies that constructing β^* as accurately as possible, which mean producing a representation of β^* that can be stored compactly than its original representation.
Many classes of signals exhibit sparsity when transformed into the appropriate basis, such as a wavelet basis. Such transform can be represented as an orthonormal matrix $\Psi \in \mathbb{R}^{d \times d}$, constructed so that

$$\theta^* := \Psi' \beta^* \in \mathbb{R}^d$$

corresponds to the vector of transformed coefficients. If θ^* is known to be sparse then only a fraction of the coefficients, say the $s < d$ largest coefficients in absolute value can be retained.
In the transformed space, the model takes the form

$$y = \theta^* + \varepsilon$$

where $y = \Psi' \tilde{y}$, and $\Psi' \tilde{\varepsilon}$. If $\tilde{\varepsilon} \sim N(0, \sigma^2)$, then it is invariant under orthogonal transformation and the original and transformed observations $\tilde{y}$ and y are examples of Gaussian sequence models touched in the earlier example, both with $n = d$. If θ^* is known to be sparse then it is natural to consider estimators based on thresholding. Wainwright (2019) shows that for a hard threshold of $\lambda > 0$, we may have hard-threshold or soft-threshold estimates of θ^*.

Example (Lifting and non-linear functions): Consider the n pair of observations $\{(y_i, t_i)\}_{i=1}^n$, where each pair is lined via the model

$$y_i = f(t_i; \theta) + \varepsilon_i,$$

where

$$f(t_i; \theta) = \theta_1 + \theta_2 t_i + \theta_3 t_i^2 + \cdots + \theta_{k+1} t_i^k.$$

This non-linear problem can be converted into an instance of linear regression model, by defining the $n \times (k+1)$ matrix

$$X = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^k \end{bmatrix}$$

which once again leads to the general linear model

$$y = X\theta + \varepsilon.$$

If we were to extend the univariate function above to a multivariate functions in D dimensions, there are $\binom{k}{D}$ possible multinomials of degree k in dimension D. This leads to an exponentially growing model with dimension of the magnitude D^k, so that the sparsity assumptions become essential.

To build intuition, we start we the simplest case were there observations are noiseless. Essentially, we wish to find a solution θ to the linear system

$$y = X\theta,$$

where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times d}$, such that $d > n$. When $d > n$, this is an undetermined set of linear equations, so there is a whole subspace of solutions.
If we have a sparse solution that means that there is a vector $\theta^* \in \mathbb{R}^d$, with at most $s \ll d$ non-zero entries and such that $y = X\theta^*$.
The goal is to find this sparse solution to the linear system.

This problem can be expressed as a non-convex optimization problem involving the l_0-"norm".

Question: The l_0-norm has been put in quotation marks, as it is not considered a proper norm. Why is that?

Let us define

$$\|\theta\|_0 := \sum_{j=1}^{d} \mathbb{1}[\theta_j \neq 0]$$

where $\mathbb{1}$ is an indicator function. Thus, the optimization problem is

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \quad \text{such that} \quad X\theta = y$$

Solving this leads to obtaining a solution to the linear equations that has the fewest number of non-zero entries. How can we solve the above optimization problem? The constraint set is simply a subspace, but the cost function is non-differentiable and non-convex.

Algorithm for solving the l_0 optimization problem

1) For each subset $S \subset \{1, \cdots, d\}$, we form the matrix $X_S \in \mathbb{R}^{|S|}$, consisting of the columns of X indexed by S.
2) Examine the linear system $y = X_S\theta$ to see whether it has a solution $\theta \in \mathbb{R}^{|S|}$.
3) Iterate over subsets in increasing cardinality, then the first solution found would be the sparsest solution.

What would be the computational cost of this optimisation approach be? If the sparsest solution contained s non-zero entries, then we would have to search over at least

$$\sum_{j=1}^{s-1} \binom{d}{j}$$

subsets before finding it.

The next solution is to replace l_0 with the nearest convex member of the l_q family, namely the l_1 norm.

Definition (Convex relaxation)

When a non-convex optimization problem is approximated by a convex programme.

In this setting this leads to the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{such that} \quad X\theta = y.$$

The constraint sex is a subspace (hence convex), and the cost function is piecewise linear and thus convex as well. The l_1 optimisation problem is a linear programme, since any piecewise linear convex cost can always be reformulated as the maximum of a collection of linear functions. The above optimisation problem is referred to as basis pursuit linear programme.

When is solving the basis pursuit problem

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{such that} \quad X\theta = y.$$

equivalent to solving the l_0 problem below?

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \quad \text{such that} \quad X\theta = y$$

Suppose $\theta^* = \mathbb{R}^d$ such that $y = X\theta^*$. Moreover, the vector θ^* has the support $S \subset \{1, 2, \cdots, d\}$, which means that $\theta_j^* = 0$ for all $j \in S^C$. The success of the basis pursuit should depend on how the nullspace of X is related to this support, where by definition

$$\text{null}(X) := \{\Delta \in \mathbb{R}^d \mid X\Delta = 0\}.$$

Since $X\theta^* = y$ by assumption, any vector of the form $\theta^* + \Delta$ for some $\Delta \in \text{null}(X)$ is feasible for the basis pursuit programme.

References

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.