Recovery in the noiseless setting

Kaveh S. Nobari

Lectures in High-Dimensional Statistics

Department of Mathematics and Statistics
Lancaster University

Contents

To build intuition, we start with the simplest case, where the 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.

1. Recovery in noiseless setting
   - l_1-based relaxation
   - Exact recovery and restricted null space
   - Sufficient conditions for restricted null space

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 set 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 [see Chen and Donoho (1998)].

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.

Now let us consider the tangent cone of the l_1-ball at θ^*, given by

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

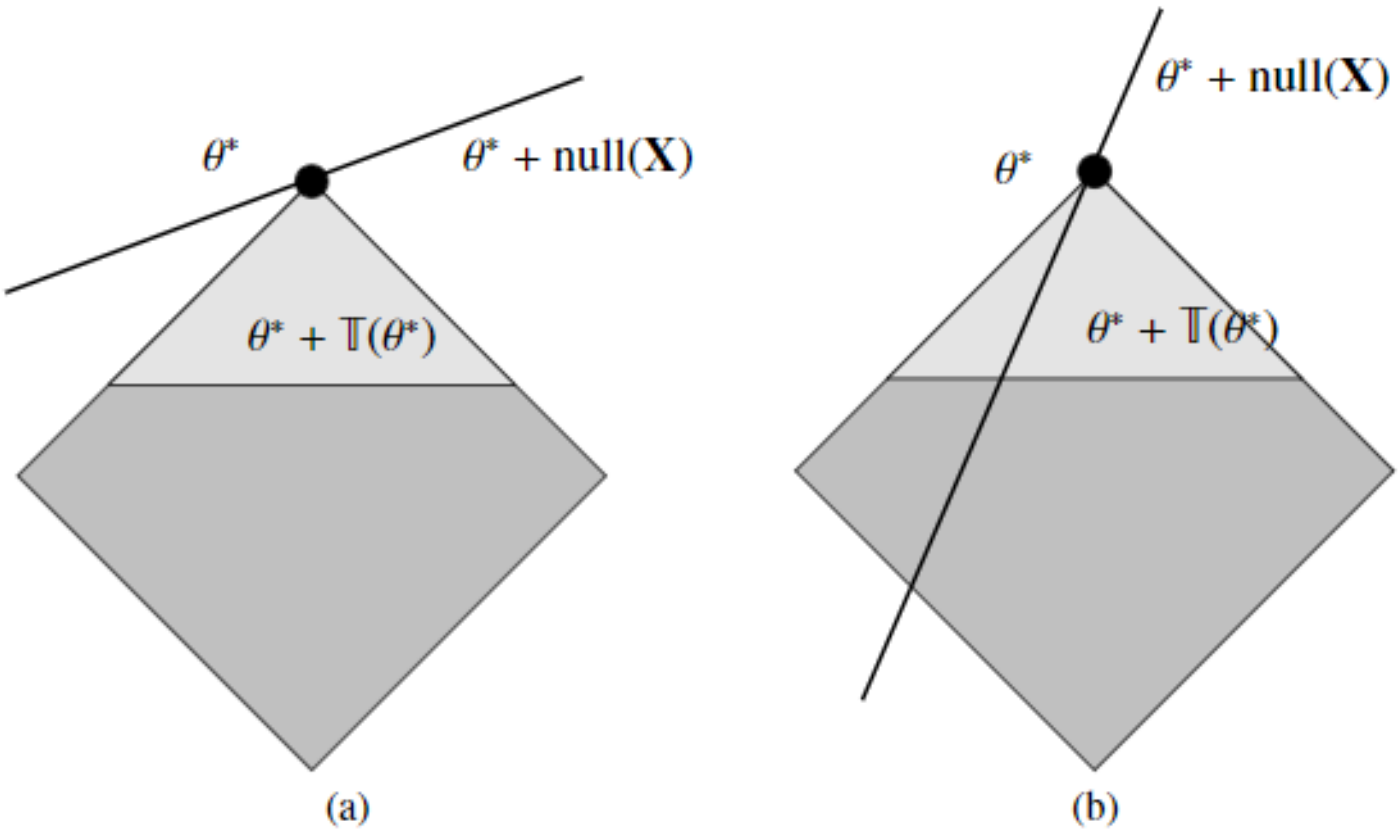

Figure 7.2 Geometry of the tangent cone and restricted nullspace property in $d = 2$ dimensions. (a) The favorable case in which the set $\theta^* + \text{null}(\mathbf{X})$ intersects the tangent cone only at θ^*. (b) The unfavorable setting in which the set $\theta^* + \text{null}(\mathbf{X})$ passes directly through the tangent cone.

- The set $\mathbb{T}(\theta^*)$ captures the set of all directions relative to θ^* along which the l_1-norm remains constant or decreases.
- The solid line $\theta^* + \text{null}(X)$ corresponds to the set of all vectors that are feasible for the basis pursuit linear programme, in the sense that $X(\theta^* + \text{null}(X)) = y$.
- In the above figures, if θ^* is optimal, then the tanget line $\theta^* + \text{null}(X)$ must only intersect with the tangent cone at θ^* implying that $\text{null}(X)$ at this point is zero vector.
- This intution leads to a condition on X, known as the restricted nullspace property.

Let us define the cone subset

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

which corresponds to the cone of vectors whose l_1-norm off the support (i.e. S^c), is dominated by the l_1-norm on the support (i.e. S). Using the defined cone subset, we can now formally define the restricted nullspace property

Definition (Restricted nullspace property)

The matrix X is said to satisfy the restricted nullspace property with respect to S if $\mathbb{C}(S) \cap \text{null}(X) = 0$.

Let us now consider an alternative way of capturing the behavior of the tangent cone $\mathbb{T}(\theta^*)$ that is independent of θ^*, one which establishes that for any S-sparse vector θ^*, the tangent cone $\mathbb{T}(\theta^*)$ is contained within $\mathbb{C}(S)$, and conversely, that $\mathbb{C}(S)$ is contained in the union of such tangent cones.

More precisely, the restricted null space property is equivalent to the success of basis pursuit in the following sense:

Theorem

(a) *The matrix X satisfies the restricted null space property with respect to S.*

(b) *For any vector $\theta^* \in \mathbb{R}^d$,with support S, the basis pursuit programme applied with $y = X\theta$ has a unique solution - i.e., $\hat{\theta} = \theta^*$.*

Proof:

We begin by proving (a): since both $\hat{\theta}$ and θ^* are feasible for the basis pursuit programme, and since $\hat{\theta}$ is optimal, we have $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$. Defining the error vector $\hat{\Delta}$ as follows

$$\hat{\Delta} := \hat{\theta} - \theta^*.$$

By construction we have $X\theta^* = X\hat{\theta} = X(\theta^* + \hat{\Delta})$. From here we have

$$\begin{aligned}\|\theta^*_S\|_1 = \|\theta^*\|_1 &\geq \|\hat{\theta}\|_1 \\ &\geq \|\theta^* + \hat{\Delta}\|_1 \\ &= \|\theta^* + \hat{\Delta}_S + \hat{\Delta}_{S^c}\|_1 \\ &= \|\theta^*_S + \underbrace{\theta^*_{S^c}}_{=0} + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1\end{aligned}$$

From the triangle inequality, we know that

$$\|\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 \leq \|\theta^*_S + \hat{\Delta}_S\|_1 \leq \|\theta^*_S\|_1 + \|\hat{\Delta}_S\|_1$$

therefore

$$\begin{aligned}\|\theta^*_S\|_1 = \|\theta^*\|_1 &\geq \|\theta^* + \hat{\Delta}\|_1 \\ &= \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\ &\geq \|\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1\end{aligned}$$

rearranging the above yields

$$\begin{aligned} \|\theta^*\|_1 &\geq \|\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\ \|\theta^*\|_1 - \|\theta^*_S\|_1 &\geq -\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\ 0 &\geq -\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\ \|\hat{\Delta}_S\|_1 &\geq \|\hat{\Delta}_{S^c}\|_1 \end{aligned}$$

which proves that $\hat{\Delta} \in \mathbb{C}(S)$. However, by construction, $X\hat{\Delta} = 0$, which means that $\hat{\Delta} \in \text{null}(X)$ too. By the assumption imposed earlier, this implies that $\hat{\Delta} = 0$ or that $\hat{\theta} = \theta^*$

For (b) it suffices to show that, if the l_1 relaxation succeeds for all S-sparse vectors, then the set $\text{null}(X)\backslash\{0\}$ has no intersection with $\mathbb{C}(S)$.

For a given vector $\theta^* \in \text{null}(X)\backslash\{0\}$ consider the basis pursuit problem

$$\min_{\beta \in \mathbb{R}^d} \|\beta\|_1, \quad \text{such that} \quad X\beta = X \begin{bmatrix} \theta_S * \\ 0 \end{bmatrix}$$

. By assumption, the unique optimal solution will be $\hat{\beta} = (\theta^*_S, 0)'$. Since $X\theta^* = 0$ by assumption, the vector $(0, -\theta_{S^c})'$ is also feasible for the problem, and, by uniqueness, we must have $\|\theta^*_S\|_1 < \|\theta^*_{S^c}\|_1$, implying that $\theta \notin \mathbb{C}(S)$ as claimed.

- To ensure that for any vector $\mathbb{R}^d$ with support S, the basis pursuit programme applied with $y = X\theta^*$ has unique solution $\hat{\theta} = \theta^*$, the matrix X has to satisfy the restricted nullspace property.
- The earliest sufficient conditions were based on the incoherence parameter of the design matrix, which is the quantity

$$\delta_{pw}(X) = \max_{j,k=1,\cdots,d} \left| \frac{\langle X_j, X_k \rangle}{n} - \mathbb{1}\{j = k\} \right|$$

  where X_j and X_k are the k^{th} and j^{th} columns of the matrix X respectively and $\mathbb{1}\{.\}$ denotes an indicator function.
- X is rescaled by dividing by $\sqrt{n}$, then $X_j' X_j = 1$, which makes is more readily interpretable. The parameter $\delta_{pw}(X)$ essentially defines the maximum absolute value of cross-correlations between the columns of X.

In what follows, through Exercise 7.3 of Wainright (2019), it will be shown that a small mutual (pairwise) incoherence is sufficient to guarantee a uniform version of the restricted nullspace property.

Proposition

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.

Proof:

Choose a vector θ such that $X\theta = 0$. For some set S subject to $|S| \leq s$, we have $\theta = \theta_S + \theta_{S^c}$,and $X(\theta_S + \theta_{S^c}) = 0$. Thus, $X\theta_S = -X\theta_{S^c}$. Let us lower bound the l_2-norm of the left hand side of the former equation

$$\begin{aligned}\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \frac{(X\theta_S)' X\theta_S}{n} = \frac{\theta_S' X' X \theta_S}{n} \\ &= \frac{\theta_S' X' X \theta_S}{n} - \theta_S'\theta_S + \theta_S'\theta_S \\ &= \theta_S' \left(\frac{X'X}{n} - I \right) \theta_S + \|\theta_S\|_2^2\end{aligned}$$

since we have the inequality

$$u' M v \leq |||M|||_2 \|u\|_1 \|v\|_1$$

The term $\theta_S' \left(\frac{X'X}{n} - I \right) \theta_S$ can be expressed as follows

$$\theta_S' \left(\frac{X'X}{n} - I \right) \theta_S \leq \left| \left| \left| \frac{X'X}{n} - I \right| \right| \right|_2 \|\theta_S\|_1 \|\theta_S\|_1$$

Thus,

$$\begin{aligned} \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S' \left(\frac{X'X}{n} - I \right) \theta_S + \|\theta_S\|_2^2 \\ &\geq -\left| \left| \left| \frac{X'X}{n} - I \right| \right| \right|_2 \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \end{aligned}$$

since the mutual incoherence parameter is the smallest constant $\delta_{pw}(X)$ such that

$$\left| \left| \left| \frac{X'X}{n} - I \right| \right| \right|_2 \leq \delta_{pw}(X)$$

we would thus have

$$\begin{aligned}\left\|\frac{X\theta_S}{\sqrt{n}}\right\|_2^2 &= \theta_S'\left(\frac{X'X}{n}-I\right)\theta_S+\|\theta_S\|_2^2\\ &\geq -\left|\left|\left|\frac{X'X}{n}-I\right|\right|\right|_2\|\theta_S\|_1^2+\|\theta_S\|_2^2\\ &\geq -\delta\|\theta_S\|_1^2+\|\theta_S\|_2^2\end{aligned}$$

Moreover, we have the inequality

$$\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2$$

which leads to

$$\left\|\frac{X\theta_S}{\sqrt{n}}\right\|_2^2 = \theta_S'\left(\frac{X'X}{n}-I\right)\theta_S+\|\theta_S\|_2^2 \tag{1}$$

$$\geq -\left|\left|\left|\frac{X'X}{n}-I\right|\right|\right|_2\|\theta_S\|_1^2+\|\theta_S\|_2^2 \tag{2}$$

$$\geq -\delta\|\theta_S\|_1^2 + \|\theta_S\|_2^2 \tag{3}$$

$$\geq -\delta s\|\theta_S\|_2^2 + \|\theta_S\|_2^2 = (1-\delta s)\|\theta_S\|_2^2 \tag{4}$$

Since $X\theta_S = -X\theta_{S^c}$ we would also have

$$\left\|\frac{X\theta_S}{\sqrt{n}}\right\|_2^2 = \left|\left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle\right| \tag{5}$$

$$= \left|\theta_S'\left(\frac{X'X}{n} - I\right)\theta_S + \underbrace{\theta_S'\theta_{S^c}}_{=0}\right| \tag{6}$$

$$\leq \delta\|\theta_S\|_1\|\theta_{S^c}\|_1 \tag{7}$$

$$\leq \delta\sqrt{s}\|\theta_S\|_2\|\theta_{S^c}\|_1 \tag{8}$$

Relating equations (1) and (5), we have

$$(1-\delta s)\|\theta_S\|_2^2 \leq \left\|\frac{X\theta_S}{\sqrt{n}}\right\|_2^2 \leq \delta\sqrt{s}\|\theta_S\|_2\|\theta_{S^c}\|_1$$

Hence, we may write the above as

$$(1-\delta s)\|\theta_S\|_2^2 \leq \delta\sqrt{s}\|\theta_S\|_2\|\theta_{S^c}\|_1 \tag{9}$$

$$\|\theta_S\|_2^2 \leq \frac{\delta\sqrt{s}}{(1-\delta s)}\|\theta_S\|_2\|\theta_{S^c}\|_1 \tag{10}$$

Recall the inequality $\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2$. Thus, multiplying equation (10) by $\sqrt{s}$, we will have

$$\sqrt{s}\|\theta_S\|_2^2 \leq \frac{s\delta}{(1-\delta s)}\|\theta_S\|_2\|\theta_{S^c}\|_1$$

$$\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2 \leq \frac{s\delta}{(1-\delta s)}\|\theta_S\|_2\|\theta_{S^c}\|_1$$

$$\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2 \leq \frac{s\delta}{(1-\delta s)}\|\theta_{S^c}\|_1$$

Assuming $\delta \leq \frac{1}{2s}$, then $\|\theta_S\|_1 \leq \|\theta_{S^c}\|_1$, therefore the restricted nullspace property holds.

A more related but sophisticated sufficient condition is the Restricted Isometry Property (RIP). This can be understood as a generalisation of the pairwise incoherence condition, based on looking at conditioning of larger subsets of columns.

Definition (Restricted isometry property)

For a given integer $s \in \{1, \cdots, d\}$, we say that $X \in \mathbb{R}^{n \times d}$ satisfies the RIP of order s with constant $\delta_s(X) > 0$, if

$$\left|\left|\left| \frac{X_S' X_S}{n} - I_s \right|\right|\right|_2 \leq \delta_s(X) \quad \text{for all subests } S \text{ of size at most } s$$

For $s = 1$, we would have

$$\left|\left|\left| \frac{X_j' X_j}{n} - 1 \right|\right|\right|_2 \leq \delta_1$$

$$\left| \frac{\|X_j\|_2^2}{n} - 1 \right| \leq \delta_1$$

which implies

$$1 - \delta_1 \leq \frac{\|X_j\|_2^2}{n} \leq 1 + \delta_1$$

for all $j = 1, \cdots, d$. Now consider $s = 2$, and suppose the matrix $X/\sqrt{n}$ has unit-norm columns. Then we would have

$$\frac{X'_{\{j,k\}} X_{\{j,k\}}}{n} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{\|X_j\|_2^2}{n} - 1 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & \frac{\|X_k\|_2^2}{n} - 1 \end{bmatrix} = \begin{bmatrix} 0 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & 0 \end{bmatrix}$$

Now let us consider the l_2-matrix norm, which is the maximum singular value - i.e.,

$$\left\| \frac{X'_{\{j,k\}} X_{\{j,k\}}}{n} - I_2 \right\|_2 = \max_{j \neq k} \left| \frac{\langle X_j, X_k \rangle}{n} \right| = \delta_{pw}(X)$$

Definition (Sandwich relation)

For any matrix X and sparsity level $s \in 2, \cdots, d$, we have the sandwich relation

$$\delta_{pw}(X) \leq \delta_s(X) \leq s\delta_{pw}(X)$$

and neither bound can be improved in general.

Although RIP imposes constraints on much larger submatrices than pairwise incoherence, the magnitude of the constraints required to guarantee uniform RNS property can be milder. Suitable control on the RIP constants implies that the RNS property holds:

Proposition

If the RIP constant of order $2s$ is bounded as $\delta_{2s}(X) < 1/2$, then the uniform RNS holds for any subset S of cardinality $|S| \leq s$.

Like pairwise incoherence constant, control on the RIP constants is sufficient condition for the BPLP to succeed. A major advantage of the RIP is that for various classes of random design matrices, it can be used to

guarantee exactness of basis pursuit using a sample size n that is much smaller than that guaranteed by pairwise incoherence. The RIP approach overcomes the "quadratic barrier" - i.e., the requirement that the sample size n scales quadratically in the sparsity s, as in the pairwise incoherence approach.

References

Chen, S. S. and Donoho, D. L. (1998). Application of basis pursuit in spectrum estimation. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 3, pages 1865–1868. IEEE.