Estimation in High Dimensions: A Geometric Perspective. Part 1

Jiapeng Liu

Department of Statistics
Purdue University

April 20, 2016

(Work by Roman Vershynin)

Introduction

High dimensional estimation problems:

Parameters: $x \in \mathbb{R}^n$. Unknown.

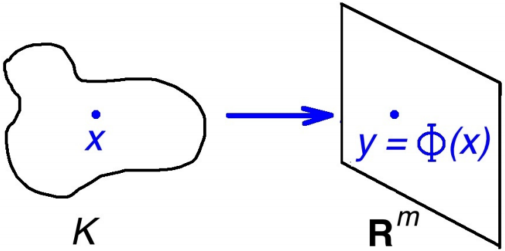
Map: $\Phi : \mathbb{R}^n \longrightarrow \mathbb{R}$. Known.

Measurement vector: $y = \Phi(x) \in \mathbb{R}^m$. Known.

Goal: Estimate x from y.

Introduction

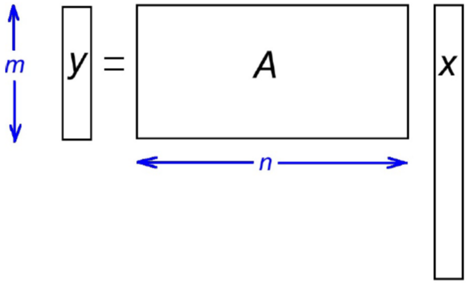
Prior information (model): $x \in \mathbf{K}$, where $\mathbf{K} \subset \mathbb{R}^n$ is a known **feasible set**.



Example of K:
$\{$s-sparse vectors in $\mathbb{R}^n\}$;
$\{$low-rank matrix in $\mathbb{R}^{n_1 \times n_2}\}$
etc.

Introduction

Linear Model : $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a $m \times n$ matrix.



Introduction

- In principle, it should be possible to estimate $\mathbf{x}$ from $\mathbf{y}$ with

$$m = O(n)$$

observations.

- If K happens to be $low - dimensional$, with algebraic dimension $dim(K) = d \ll n$. Then in this case, the estimation should be possible with fewer observations.

$$m = O(d) = o(n).$$

- However, it rarely happens that feasible sets of interest have small algebraic dimension.
For example, {s-sparse vectors in $\mathbb{R}^n$} has full dimension n but tend to have $low - complexity$.

Introduction

Question

How to estimate $\mathbf{x}$ when K is $high-dimensional$ while has $low\ complexity$?

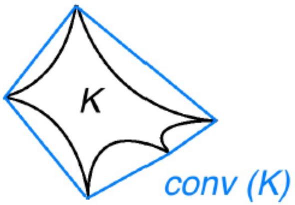
Goal :

- 1. develop geometric intuition about high dimensional sets;
- 2. explain results of asymptotic convex geometry which validate this intuition;
- 3. demonstrate connections between high dimensional geometry and high dimensional estimation problems.

High dimensional convex geometry

Convexity

The set K may be non-convex. Then convexity: $K \longmapsto \operatorname{conv}(K)$.



Question: What do convex sets look like in high dimensions?

This is the main question of **Asymptotic Convex Geometry = Geometric Functional Analysis**.

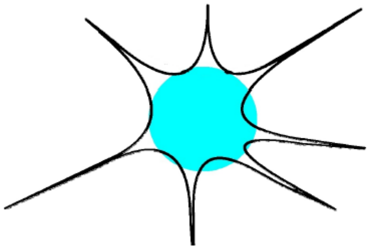
High dimensional convex geometry

Intuition about convex sets

Main message of asymptotic convex geometry:

$$K \approx Bulk + Outliers.$$

$Bulk$ = small diameter, round ball, makes up most volume of K.
$Outliers$ = few long tentacles, contain little volume..



V. Milman's heuristic picture of a convex body in high dimensions.

Example: The l_1 and l_2 ball

$\mathbf{l_1}$ ball :
The volume of a unit l_1 ball is

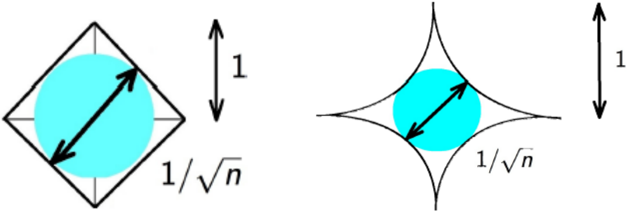
$$Vol_n(B_1^n) = \frac{2^n}{n!}, \qquad Vol_n(B_1^n)^{1/n} \approx \frac{2e}{n}$$

$\mathbf{l_2}$ ball :
The volume of a unit l_2 ball is

$$Vol_n(B_2^n) = \left(\sqrt{\frac{2\pi e}{n}}\right)^n, \qquad Vol_n(B_2^n)^{1/n} = \sqrt{\frac{2\pi e}{n}}$$

Example: The l_1 ball

$K = \text{conv}(\pm e_i) = \{ x : \|x\|_1 \leq 1 \} = B_1^n$ is the unit l_1 ball.



The left figure is the standard figure, but the right figure is more accurate.

$$Vol(K)^{1/n} \asymp Vol(B)^{1/n} \asymp \frac{1}{n}$$

Here B is the Euclidean ball in K, with diameter $2/\sqrt{n}$.

Rigorous results for this intuition

Concentration of volume

Definition (Isotropic)

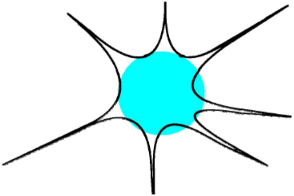
A set K is *isotropic* if the random vector X distributed uniformly in K (according to the Lebesgue measure) and satisfy:

$$\mathbb{E}X = 0, \quad \mathbb{E}XX^T = I_n.$$

- Isotropy means the geometry on the manifold is the same regardless of direction;
- Isotropy is just an assumption of proper scaling – one can always make a convex body K isotropic by applying a suitable invertible linear transformation.
- With this scaling, most of the volume of K is located around the Euclidean sphere of radius $\sqrt{n}$.

Rigorous results for this intuition

Concentration of volume



Isotropy assumption: X $\sim$ Unif(K) satisfies $\mathbb{E}X = 0$, $\mathbb{E}XX^T = I_n$.
$\mathbb{E}\|X\|_2^2 = n$, so the **radius** of that ball is $\sqrt{n}$.

Theorem (concentration of volume) [Paouris '06]

$$\mathbb{P}\{\|X\|_2 > t\sqrt{n}\} \le \exp(-ct\sqrt{n}),\ \ t \ge 1.$$

Theorem (thin shell) [Klartag '07]

$$\mathbb{P}\{\big|\|X\|_2 - \sqrt{n}\big| > \epsilon\sqrt{n}\} \le C \exp\left(-c\epsilon^3 n^{1/2}\right).\quad \forall\ \epsilon \in (0,1)$$

C and c denote positive absolute constants.

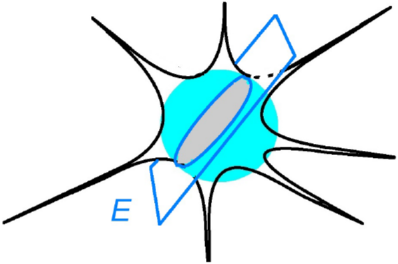
Low dimensional random sections

Question

What the *random sections* of a high dimensional convex set K look like?

Suppose E is a random subspace of $\mathbb{R}^n$ with fixed dimension d.

If d is sufficiently small and the bulk of K is a round ball, then we should expect the section $K \cap E$ to be a round ball as well.
i.e. E **misses the outliers**, **passes through the bulk** of K.



Low dimensional random sections

Theorem (Dvoretzky's theorem)

Let K be an $origin - symmetric$ convex body in $\mathbb{R}^n$ such that the ellipsoid of maximal volume contained in K is the unit Euclidean ball B_2^n. Fix $\epsilon \in (0, 1)$. Let E be a random subspace of dimension $d = c\epsilon^{-2} log\ n$ drawn from the Grassmanian $G_{n,d}$ according to the Haar measure. Then there exists $R \geq 0$ such that with high probability (say, 0.99) we have

$$(1 - \epsilon)B(R) \subseteq K \cap E \subseteq (1 + \epsilon)B(R)$$

Here B(R) is the centered Euclidean ball of radius R in the subspace E.

High dimensional random sections

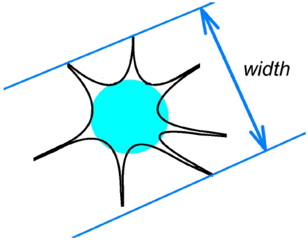
Question

What the $random\ sections$ of a high dimensional convex set K look like when d is large?

- If d is large, we can no longer expect such sections to be round.
- As the $codimension$ decreases, the random subspace E becomes larger and it will probably pick more and more of the outliers (tentacles) of K.
- The shape of such sections $K \cap E$ is difficult to describe.
- Nevertheless, we can accurately predict the $diameter$ of $K \cap E$. A bound on the $diameter$ is known in Asymptotic Convex Geometry as the low M^* estimate, or M^* bound.

Mean width

Before we move on, we need a new concept called $Mean\ Width$.



Definition(Gaussian Mean Width)

$$w(K) := \mathbb{E}\ \sup_{x \in K-K} \langle g, x \rangle, \text{ where } g \sim N(0, I_n),$$

K is a bounded set and $K - K = \{\mathbf{u} - \mathbf{v} : \mathbf{u} - \mathbf{v} \in K\}$.

Mean width

Remark :

- The concept of mean width captures important geometric characteristics of sets in $\mathbb{R}^n$.
- One can mentally place it in the same category as other classical geometric quantities like volume and surface area.
- By the χ^2 distribution, $\mathbb{E}\|g\|_2 \asymp \sqrt{n}$, so
$$w(K) \approx \sqrt{n} \cdot \text{width of K in a random direction}$$
- The mean width is invariant under translations, orthogonal transformations, and taking convex hulls, especially:
$$w(conv(K)) = w(K)$$

Example: Some useful mean width

Unit $\mathbf{l_2}$ ball :

$$w(K) = 2\mathbb{E}\|g\|_2 \asymp \sqrt{n}$$

$\mathbf{s-sparse\ vectors}$: Let K consist of all unit s-sparse vectors in $\mathbb{R}^n$ those with at most s non-zero coordinates:

$$K = \{x \in \mathbb{R}^n : \|x\|_2 = 1; \|x\|_0 \leq s\} :$$

Then,

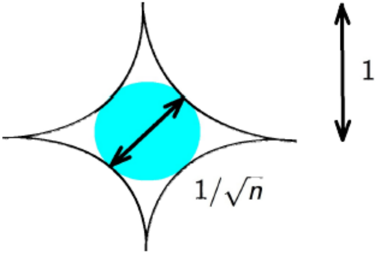
$$w(K) \asymp \sqrt{s\ log(2n/s)}$$

Example:Some useful mean width

Unit $\mathbf{l_1}$ ball :

Mean Width: $w(K) := \mathbb{E}\ \sup_{x \in K-K} \langle g, x \rangle$

$\asymp \sqrt{n} \cdot \mathbb{E}$[width of K in random direction].



w(K) $\asymp \sqrt{logn}$, w(B) $\asymp$ 1.

Hence Mean Width **sees the bulk**, ignores the outliers.

Computing Mean Width algorithmically

Question

How to estimate of a given set K fast and accurately?

- Gaussian concentration of measure implies that, with high probability, the random variable

$$w(K, g) = \sup_{x \in K-K} \langle g, x \rangle$$

  is close to its expectation $w(K)$.
- It is enough to generate a single realization of a random vector $g \sim N(0, I_n)$ and compute $w(K, g)$, this should produce a good estimator of $w(K)$.
- Since we can convexify K without changing the mean width, computing this estimator is a *convex optimization problem* (and often even a linear problem if K is a polytope).

Computing Mean Width theoretically

Definition (Metric Entropy)

Let N(K,t) denote the smallest number of Euclidean balls of radius t whose union covers K. Usually N(K,t) is referred to as a *covering number* of K, and logN(K,t) is called the *metric entropy*.

The mean width is related to the *metric entropy*.

Theorem (Sudakov's and Dudley's inequities)

For any bounded subset K of $\mathbb{R}^n$, we have

$$c \ \mathrm{sup}_{t>0} t\sqrt{logN(K,t)} \leq w(K) \leq C \int_0^\infty \sqrt{logN(K,t)}dt.$$

- The lower bound is Sudakov's inequality and the upper bound is Dudley's inequality.
- Neither Sudakov's nor Dudley's inequality are tight for all sets K.

Random sections of small codimension: M^* bound

Theorem (M^* estimate) [Mendelson-Pajor-Tomczak '07]

Let K be a bounded subset of $\mathbb{R}^n$. Let E be a random subspace of $\mathbb{R}^n$ of a fixed codimension m, drawn from the Grassmanian $G_{n,n-m}$ according to the Haar measure. Then

$$\mathbb{E}\,\mathrm{diam}(K \cap E) \leq \frac{Cw(K)}{\sqrt{m}}.$$

Here $w(K)$ is the **mean width** of K.

- For subspaces E of not very high dimension, where $m = O(n)$, the M^* bound states that the size of the random section $K \cap E$ is bounded by the spherical mean width of K.
- When the dimension of the subspace E grows toward n (so the codimension m becomes small), the diameter of $K \cap E$ also grows by a factor of $\sqrt{n/m}$.

Application: Estimation of Linear Observations

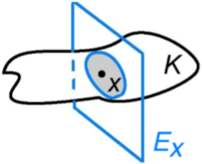
Recover x $\in$ K $\subset \mathbb{R}^n$ from $y = Ax \in \mathbb{R}^n$.

If $m \leq n$, problem is $\mathbf{ill-posed}$.

What do we know? x belongs to both K and the affine subspace

$$E_x := \{x' : Ax' = y\} = x + ker(A)$$

Both K and E_x are $\mathbf{known}$.



If $\mathrm{diam}(K \cap E_x) \leq \epsilon$, then x can be recovered with error ϵ.

Estimation of Linear Observations

When $\mathrm{diam}(K \cap E_x) \leq \epsilon$?

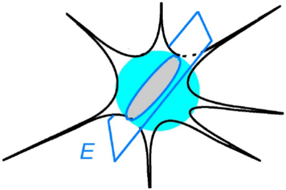
Assume K is convex and origin-symmetric, then K - K is 2K.
Let E=ker(A).

For any $x' \in K$ such that $Ax' = y$,

$$x - x' \in 2K \text{ and } x - x' \in E$$

Conclusion : If $\mathrm{diam}(K \cap E) \leq \epsilon$, then any $x \in K$ can be recovered from y=Ax with error ϵ.

Estimation of Linear Observations



Remaining question :When is $\mathrm{diam}(K \cap E) \le \epsilon$?

If A is an random matrix then E is a **random subspace**.
Another version of $\mathbf{M}^*$ **estimate** is

$$\mathrm{diam}(K \cap E) \le \frac{Cw(K)}{\sqrt{m}} \quad \text{with high probability.}$$

Equate with ϵ and obtain the **sample size** (# of measurements):

$$m \asymp \epsilon^{-2} w(K)^2$$

Estimation of Linear Observations

Conclusion

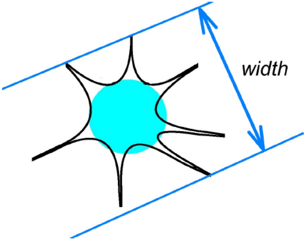
Let K be a convex and origin-symmetric set in $\mathbb{R}^n$.
Let A be an m×n *random matrix*.
If $m \asymp w(K)^2$ then one can accurately recover any signal $x \in K$ from m random linear measurements given as $y = Ax \in \mathbb{R}^m$

The recovery is done by the **convex program**

$$\text{Find } x' \in K \text{ such that } Ax' = y$$

i.e. find x' consistent with the model (K) and measurements (y).

Estimation of Linear Observations



Remark :

- 1. If the signal set K is not convex, then convexify;

$$w(conv(K)) = w(K)$$

- 2. If the signal set K is not origin-symmetric, then symmetrize;

$$w(K - K) \leq 2w(K)$$

- 3. Mean width can be efficiently estimated. Randomized linear program:

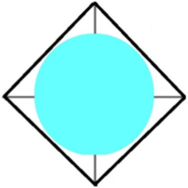
$$w(K) \approx \sup_{x \in K-K} \langle g, x \rangle, \quad g \sim N(0, I_n)$$

Information theory viewpoint

The sample size $m = w(K)^2$ is an **effective dimension** of K, the amount of information in K.

Conclusion : One can effectively recover any signal in K from $w(K)^2$ random linear measurements.

Example 1. $K = B_1^n = conv(\pm e_i)$. $w(K)^2 \sim log\ n$.
So, one can effectively recover any signal in B_1^n from $log\ n$ random linear measurements.



Information theory viewpoint

Example 2.
$K = \{\text{s} - \text{sparse unit vectors in } \mathbb{R}^n\}$. $w(K)^2 \asymp slog\ n.$

Conclusion : One can effectively recover any s-sparse signal from $slog\ n$ random linear measurements.

This is a well-known result in **compressed sensing**.
(Warning: exact recovery is not explained by this geometric reasoning.)

Remark. $log\binom{n}{s} \asymp slog\ n$. bits are required to specify the sparsity pattern.

Estimation of Noisy Linear Observations

Assume $\epsilon \geq 0$, we have:

$$y = Ax + \mu, \quad \frac{1}{m}\|\mu\|_1 = \frac{1}{m}\sum_{i=1}^{m}|\mu_i| \leq \epsilon$$

Here A is an $m \times n$ Gaussian matrix as before, the noise vector μ may be unknown and have arbitrary structure.
K is any bounded set in $\mathbb{R}^n$.

Theorem (Feasibility program)

Choose $\hat{x}$ to be any vector satisfying

$$\hat{x} \in K \quad \text{and} \quad \frac{1}{m}\|A\hat{x} - y\|_1 \leq \epsilon$$

Then

$$\mathbb{E}\,\sup_{x\in K}\|\hat{x} - x\|_2 \leq \sqrt{8\pi}\left(\frac{Cw(K)}{\sqrt{n}} + \epsilon\right).$$

Estimation of Noisy Linear Observations

Question

How to find such a x satisfy the Feasibility program?

- **Assumption** : K is a star-shaped bounded set in $\mathbb{R}^n$ with nonempty interior, which means:

$$tK \subseteq K \ for \ all \ t \in [0,1]$$

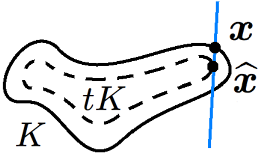
- For each point $x \in \mathbb{R}^n$, the $Minkowski \ functional$ of K, $\|x\|_K$ is defined by:

$$\|x\|_K = inf\{\lambda > 0 : \lambda^{-1}x \in K\}$$

- It is easy to see that

$$K = \{x : \|x\|_K \leq 1\}$$

Estimation of Noisy Linear Observations



Theorem (Optimization program)

Choose $\hat{x}$ to be a solution to the program

$$minimize \ \|\hat{x}\|_K \quad subject \ to \ \ \frac{1}{m}\|A\hat{x} - y\|_1 \leq \epsilon$$

Then

$$\mathbb{E} \sup_{x \in K} \|\hat{x} - x\|_2 \leq \sqrt{8\pi}\left(\frac{Cw(K)}{\sqrt{n}} + \epsilon\right).$$

References

- K. Ball (1997), An elementary introduction to modern convex geometry. *Math. Sci. Res. Inst. Publ.,*, **31**, Cambridge Univ. Press, Cambridge 1–58.
- S. Mendelson, A. Pajor, N. Tomczak-Jaegermann,(2007) Reconstruction and subgaussian operators in asymptotic geometric analysis, *Geom. Funct. Anal.* , **17**, 1248–1282.
- B. Klartag,(2007) Power-law estimates for the central limit theorem for convex sets, *J. Funct. Anal.* , **245**, 284–310.
- G. Paouris,(2006) Concentration of mass on convex bodies, *Geom. Funct. Anal* , **16**, 1021–1049.