SCIENCES SORBONNE UNIVERSITÉ

Low-rank matrix recovery

Claire Boyer

Install CVXPY for the TP

An extension of compressive sensing

- Rather than recovering a sparse vector $x \in \mathbb{C}^d$, we now aim at recovering a matrix $X \in \mathbb{C}^{d_1 \times d_2}$ from incomplete information.
- Sparsity is replaced by the assumption that X is low-rank. Indeed, the small complexity of the set of matrices with a given low rank compared to the set of all matrices makes the recovery of such matrices plausible.
- For a linear map $\mathcal{A} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^m$ with $m < d_1 \times d_2$, suppose that we are given the measurement vector

$$y = \mathcal{A}(X) \in \mathbb{C}^m$$

- The task is to reconstruct X from y. To stand a chance of success, we assume that X has rank at most

$$r \ll \min(d_1, d_2).$$

.

Rank minimization

The naive approach of solving the optimization problem

$$\min \operatorname{rank}(X) \qquad \text{s.t.} \quad \mathcal{A}(X) = y$$

is NP-hard.

Heuristics have been derived to circumvent this issue, they can be categorized into 2 classes:

1. Approximate the rank function with some surrogate functions
   1.1 Nuclear norm minimization
   1.2 Log-det heuristic
2. Solving a sequence of rank-constrained feasibility problems
   2.1 Matrix factorization based method
   2.2 Rank constraint via convex iteration

In this course, we are going to focus on 1.1.

Summary

SVD

While the concept of eigenvalues and eigenvectors applies only to square matrices, every (possibly rectangular) matrix possesses a singular value decomposition.

Proposition

Let $A \in \mathbb{C}^{m\times d}$, there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{d\times d}$, and uniquely defined nonnegative numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,d)} \geq 0$ called singular values of A such that

$$A = U\Sigma V^* \qquad \text{with} \qquad \Sigma = diag(\sigma_1, \ldots, \sigma_{\min(d,m)}) \in \mathbb{R}^{m\times d}$$

- Writing $U = [u_1| \ldots |u_m]$ and $V = [v_1| \ldots |v_d]$,
  - the vectors u_ℓ are called left singular vectors,
  - the v_ℓ are called right singular vectors.

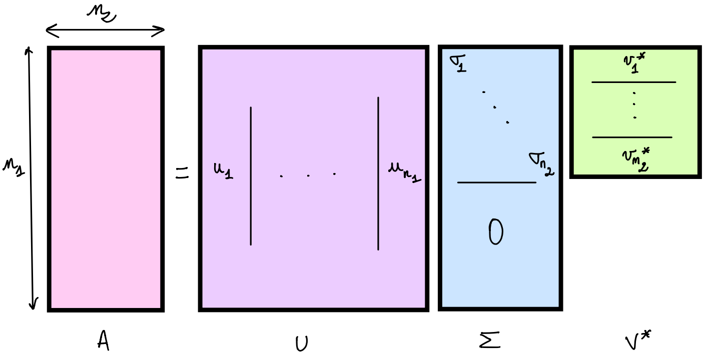

Figure: Example of SVD

- The largest and smallest singular values satisfy

$$\sigma_{\max}(A) = \sigma_1(A) = \|A\|_{2\to 2} = \max_{\|x\|_2=1} \|Ax\|_2,$$

$$\sigma_{\min}(A) = \sigma_{m\wedge d}(A) = \min_{\|x\|_2=1} \|Ax\|_2.$$

- If A has rank r, then its r largest singular values $\sigma_1 \geq \ldots \geq \sigma_r > 0$ are (strictly) positive, while $\sigma_{r+1} = \sigma_{r+2} = \ldots = 0$.
- Σ can be uniquely determined if the singular values are sorted in decreasing order. However the matrices U and V are not unique.

Lemma

One has

$$\|A\|_{2\to 2} = \sqrt{\lambda_{\max}(A^*A)} = \sigma_{\max}(A),$$

with $\lambda_{\max}(A^*A)$ the largest eigenvalue of A^*A, and $\sigma_{\max}(A)$ the largest singular value of A.

Exercise

Compute the SVD of

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}$$

On the space $\mathbb{R}^{d_1 \times d_2}$ of matrices of size $d_1 \times d_2$, we define the scalar product

$$\begin{aligned}\langle A, B\rangle &= \sum_{i,j} A_{ij} B_{ij}^* = \mathrm{Tr}(B^* A) = \mathrm{Tr}(AB^*) \\ &= \mathrm{Tr}(A^T B) = \mathrm{Tr}(BA^T) \quad \text{for scalar matrices.}\end{aligned}$$

For any matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we note $\mathrm{sp}(A)$ its spectrum:

$$\mathrm{sp}(A) = (\sigma_j)_{1 \leq j \leq d_1 \wedge d_2} \qquad \sigma_1(A) \geq \ldots \geq \sigma_{d_1 \wedge d_2}(A) \geq 0.$$

Schatten norms

- Define the Schatten norms: p-norms on the spectrum

$$\|A\|_{S_p} = \left(\sum_{j=1}^{d_1 \wedge d_2} \sigma_j(A)^p \right)^{1/p},$$

  for any $A \in \mathbb{R}^{d_1 \times d_2}$ and $p \geq 1$.
- The matrix rank can be then expressed as ??

$$\operatorname{rank}(A) = \|\operatorname{sp}(A)\|_0.$$

- The Schatten 1-norm is also called the nuclear norm or the trace norm.

$$\|A\|_{S^1} = \|A\|_* = \operatorname{Tr}(\sqrt{A^*A}) = \sum_{j=1} \sigma_j(A),$$

  where $\sqrt{A^*A}$ denotes a positive semidefinite matrix such that $BB^* = A^*A$. (More precisely, since A^*A is a positive semidefinite matrix, its square-root is well defined).

A very useful inequality

- One can show that for any matrices $X, Y \in \mathbb{R}^{d_1 \times d_2}$, such that $\sigma_1(X) \geq \ldots \geq \sigma_{d_1 \wedge d_2}(X) \geq 0$ and $\sigma_1(Y) \geq \ldots \geq \sigma_{d_1 \wedge d_2}(Y) \geq 0$

$$\sum_{j=1}^{d_1 \wedge d_2} |\sigma_j(X) - \sigma_j(Y)| \leq \sum_{j=1}^{d_1 \wedge d_2} \sigma_j(X - Y),$$

$\rightsquigarrow$ leads to triangle inequality for $\| \cdot \|_{S^1}$ ($\| \cdot \|_{S^1}$ is indeed a norm).

Schatten norms

- The Schatten 2-norm is the Frobenius norm, i.e. the Hilbert norm associated to $\langle \cdot, \cdot \rangle$:

$$\|X\|_{S_2} = \|X\|_F = \sqrt{\operatorname{Tr}(X^* X)} = \sqrt{\operatorname{Tr}(X X^*)} = \left(\sum_{jk} |X_{jk}|^2 \right)^{1/2}.$$

- The Schatten ∞-norm is the operator norm previously written as $\|\cdot\|_{2\to 2}$, i.e. the operator norm from $\ell_2^{d_2}$ to $\ell_2^{d_1}$:

$$\|X\|_{S^\infty} = \|X\|_{2\to 2} = \sup_{\|x\|_2 = 1} \|Xx\|_2.$$

After identifying matrices on $\mathbb{R}^{d_1 \times d_2}$ with vectors in $\mathbb{R}^{d_1 d_2}$, the Frobenius norm can be interpreted as an ℓ_2-norm on $\mathbb{R}^{d_1 d_2}$.

- Link between $\|\cdot\|_{2\to 2} = \|\cdot\|_{S_\infty}$ and $\|\cdot\|_F = \|\cdot\|_{S^2}$:

$$\|X\|_{S^\infty} \le \|X\|_{S^2} \qquad \text{or equivalently} \qquad \|X\|_{2\to 2} \le \|X\|_F$$

that can be showed by Cauchy-Schwarz inequality: ...

Schatten norms

- $$\|Xx\|_2^2 = \sum_i \left(\sum_j X_{ij} x_j \right)^2 \leq \sum_i \left(\sum_k x_k^2 \right) \left(\sum_j X_{ij}^2 \right) = \|X\|_F^2 \|x\|_2^2$$
- Let us recall duality results
$$\|X\|_{S^q} = \sup_{\|Y\|_{S_p} \leq 1} \langle X, Y \rangle,$$
for $\frac{1}{p} + \frac{1}{q} = 1$.
- Von Neuman's inequality says that for all $X, Y \in \mathbb{R}^{d_1 \times d_2}$,
$$\langle X, Y \rangle \leq \sum_{i=1}^{d_1 \wedge d_2} \sigma_i(X) \sigma_i(Y).$$

An idea of a necessary number of measurements

Prediction

- How many needed measurements should be required at least to reconstruct an $d_1 \times d_2$ matrix of rank r ?

A lower bound

The number of degrees of freedom for a matrix of size $d_1 \times d_2$ of rank r is

$$r(d_1 + d_2 - r)$$

Some sets

For any integer p, we call

- the set of symmetric matrices

$$\mathcal{S}^p = \{X \in \mathbb{R}^{p \times p} : X^T = X\}$$

- the cone of semi-definite positive matrices (i.e. $\forall x, \langle x, Ax \rangle \geq 0$)

$$\mathcal{S}_+^p = \{X \in \mathcal{S}^p, X \succcurlyeq 0\}$$

- the cone of definite positive matrices (i.e. $\forall x, \langle x, Xx \rangle > 0$)

$$\mathcal{S}_{++}^p = \{X \in \mathcal{S}^p, X \succ 0\}.$$

- the group of orthogonal matrices

$$\mathcal{O}(p) = \{X \in \mathbb{R}^{p \times p}, X^T X = XX^T = \mathrm{Id}_p\}$$

Summary

Analogy with CS

Consider the singular value decomposition of $X \in \mathbb{R}^{d_1 \times d_2}$, i.e. for $d = \min(d_1, d_2)$,

$$X = \sum_{\ell=1}^{d} \sigma_\ell u_\ell v_\ell^*$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ are the singular values of X.
The matrix X is of rank r if and only if the vector $\sigma = \sigma(X)$ of singular values is r-sparse, i.e., $\mathrm{rank}(X) = \|\sigma\|_0 = r$.
Introduce the so-called nuclear norm as the ℓ^1-norm of the singular values, i.e.,

$$\|X\|_* = \|X\|_{S^1} = \|\sigma(X)\|_1 = \sum_{\ell=1}^{d} \sigma_\ell(X).$$

$\rightsquigarrow$ $\|\cdot\|_{S^1}$ instead of $\|\cdot\|_*$ (confusion w/ the dual of some norm $\|\cdot\|$).

Exercise

One can check that the $\|\cdot\|_{S^1}$ is indeed a norm.

Convexification of the rank

Theorem

The convex envelope of the rank function on $B_\infty = \{X \in \mathbb{R}^{d_1 \times d_2}, \|X\|_{S^\infty} \leq 1\}$ is the nuclear norm.

Since the convex envelope of the rank is the nuclear norm, it seems natural to replace the rank in the minimization problem by the nuclear norm.

Nuclear norm minimization problem

$$\min_X \|X\|_{S^1} \qquad \text{s.t.} \quad \mathcal{A}(X) = y. \tag{NNM}$$

Nuclear Norm Minimization (NNM)

Problem NNM is a convex optimization problem which can be solved efficiently, for instance, after reformulation as a semidefinite program.

SDP formulation

$$\inf_{\substack{X\in\mathbb{C}^{d_1\times d_2},\\ Y\in\mathbb{C}^{d_1\times d_1},Z\in\mathbb{C}^{d_2\times d_2}}} \quad \mathrm{Tr}(Y)+\mathrm{Tr}(Z) \quad \text{s.t.} \quad \begin{cases} \mathcal{A}(X) & = y \\ \begin{bmatrix} Y & X^* \\ X & Z \end{bmatrix} & \succeq 0 \end{cases} \tag{SDP}$$

Theorem ((NNM) $\Leftrightarrow$ (SDP))

Problems (NNM) *and* (SDP) *are equivalent:*

1. *If* $\hat{X}$ *is solution of* (NNM), *then* $(\hat{X}, \hat{Y}, \hat{Z})$ *is solution of* (SDP) *for some* $\hat{Y}$, $\hat{Z}$.
2. *If* $(\hat{X}, \hat{Y}, \hat{Z})$ *is solution of* (SDP) *then* $\hat{X}$ *is solution of* (NNM).

Proof of Theorem (NNM $\Leftrightarrow$ SDP)

In order to show Theorem "(NNM) $\Leftrightarrow$ (SDP)", we first need the following lemma.

Lemma

Let $X \in \mathbb{R}^{d_1 \times d_2}$ and $t \geq 0$. The two following assertions are equivalent:

a) $\|X\|_{S_1} \leq t$

b) *there exist $Y \in \mathcal{S}^{d_1}, Z \in \mathcal{S}^{d_2}$ such that*

$$\mathrm{Tr}(Y) + \mathrm{Tr}(Z) \leq 2t \text{ and } \begin{bmatrix} Y & X \\ X^\top & Z \end{bmatrix} \succeq 0$$

Proof of Theorem (NNM $\Leftrightarrow$ SDP) I

We first show $1 \Rightarrow 2$. If $\hat{X}$ is solution of (NNM) , then using Lemma 7 ((a) to (b) with $t = \|\hat{X}\|_{S^1}$), there exist $\hat{Y}, \hat{Z}$ such that

$$\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z}) \leq 2\left\|\hat{X}\right\|_{S_1} \text{ and } \begin{bmatrix} \hat{Y} & \hat{X} \\ \hat{X}^\top & \hat{Z} \end{bmatrix} \succeq 0.$$

Since $\mathcal{A}(\hat{X}) = y$, $(\hat{X}, \hat{Y}, \hat{Z})$ is admissible for (SDP).
Consider (X, Y, Z) to be admissible for (SDP), one has $\mathcal{A}(X) = y$ and by optimality of $\hat{X}$ for (NNM), $\left\|\hat{X}\right\|_{S_1} \leq \|X\|_{S_1}$.

$$\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z}) \leq 2\left\|\hat{X}\right\|_{S_1} \leq 2\|X\|_{S_1}.$$

Applying Lemma 7 ((b) to (a) with $t = (\mathrm{Tr}(Y) + \mathrm{Tr}(Z))/2$), one has that

$$\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z}) \leq 2\left\|\hat{X}\right\|_{S_1} \leq 2\|X\|_{S_1} \leq 2t = \mathrm{Tr}(Y) + \mathrm{Tr}(Z).$$

Proof of Theorem (NNM $\Leftrightarrow$ SDP) II

Hence, $(\hat{X}, \hat{Y}, \hat{Z})$ is solution of (SDP).
To show $2 \Rightarrow 1$, suppose that $(\hat{X}, \hat{Y}, \hat{Z})$ is solution of (SDP). One has

$$\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z}) \leq 2t \text{ and } \begin{bmatrix} \hat{Y} & \hat{X} \\ \hat{X}^\top & \hat{Z} \end{bmatrix} \succeq 0$$

for $t = (\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z}))/2$ and since $(\hat{X}, \hat{Y}, \hat{Z})$ is admissible of (SDP). Then, using Lemma 7((b) to (a)), one has

$$\left\|\hat{X}\right\|_{S_1} \leq \frac{\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z})}{2}.$$

Let X be admissible for (NNM), i.e. $\mathcal{A}(X) = y$. Using Lemma 7 for $t = \|X\|_{S^1}$, there exist Y, Z such that

Proof of Theorem (NNM $\Leftrightarrow$ SDP) III

$\mathrm{Tr}(Y) + \mathrm{Tr}(Z) \leq 2\,\|X\|_{S_1}$ and (X, Y, Z) is admissible for (SDP). In particular, by optimality of $(\hat{X}, \hat{Y}, \hat{Z})$ for (SDP),

$$\left\|\hat{X}\right\|_{S_1} \leq \frac{\mathrm{Tr}(\hat{Y}) + \mathrm{Tr}(\hat{Z})}{2} \overset{opt.}{\leq} \frac{\mathrm{Tr}(Y) + \mathrm{Tr}(Z)}{2} \leq \|X\|_{S_1}$$

and since $\mathcal{A}(\hat{X}) = y$, $\hat{X}$ is admissible for (NNM).One can conclude that $\hat{X}$ is solution of (NNM). $\square$

A general semi-definite programming problem

$$\min_{X \in \mathcal{S}^p} \langle C, X \rangle \qquad \text{s.t.} \quad \tilde{\mathcal{A}}(X) = b, \quad X \succcurlyeq 0.$$

with $C \in \mathcal{S}^p, \tilde{\mathcal{A}} : \mathcal{S}^p \to \mathbb{R}^m$ a linear map, and $b \in \mathbb{R}^m$.

The problem (SDP) can be written as follows:

$$\min \left\langle C, \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \right\rangle, \quad \text{s.t.} \quad \tilde{A}\left(\begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \right) = y, \quad \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succcurlyeq 0,$$

where the minimum is taken over all symmetric matrices of the form $\begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix}$ with $C = \mathrm{Id}_{d_1+d_2}$, $\quad \tilde{A}\left(\begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \right) = \mathcal{A}(X) \quad$ and $\quad b = y$.

Semi-definite programming II

There exists algorithms similar to those for linear programming to approximately solve SDP problems. For instance, using `CVXOPT`

- by transforming X and C as a vector of size p^2 by concatenation with $\mathrm{vec} : \mathbb{R}^{p\times p} \to \mathbb{R}^{p^2}$
- by transforming $\tilde{A}$ in a linear operator $A^T : \mathbb{R}^{p^2} \to \mathbb{R}^m$.

Then a general SDP problem can be written as

$$\min_X \langle \mathrm{vec}(C), \mathrm{vec}(X)\rangle, A^T(\mathrm{vec}(X)) = b, X \succcurlyeq 0,$$

which can be solved using

```
sol = cvxopt.solvers.conelp(-b, A, vec(C))
```

with $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times p^2}$ and $\mathrm{vec}(C) \in \mathbb{R}^{p^2}$ are `matrix`-type objects for `CVXOPT`.

Summary

Exact reconstruction for (NNM)

The goal here is

- to determine for which measurements "vectors" $A_i \in \mathbb{R}^{d_1 \times d_2}$, for $1 \le i \le m$
- a bound on m to reconstruct any matrix X of rank r from measurements $y_i = \langle X, A_i \rangle$, $1 \le i \le m$ using (NNM).

Define

$$\Sigma_r := \left\{ X \in \mathbb{R}^{d_1 \times d_2}, \mathrm{rank}(X) = r \right\}.$$

Exact reconstruction for (NNM)

Definition

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear measurement map $\mathcal{A}(X) = (\langle X, A_i \rangle)_{1 \leq i \leq m}$. $\mathcal{A}$ is said to verify the exact reconstruction property of order r, denoted by $ER(r)$, when for any matrix $X \in \Sigma_r$:

$$
\begin{aligned}
\{X\} &= \operatorname*{arg\,min}_Z \{\|Z\|_* : \mathcal{A}(Z) = \mathcal{A}(X)\} \\
&= \operatorname*{arg\,min}_Z \{\|Z\|_{S_1} : \mathcal{A}(Z) = \mathcal{A}(X)\}
\end{aligned}
$$

One can derive the same arguments than those derived for Basis Pursuit with matrix versions of NSP and RIP.

- no proof (because they are quite similar to the vector case).
- another method of proof emphasizing the role of the non-differentiation of the objective function to force "sparsity"

Summary

rNSP

Definition

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear measurement map. $\mathcal{A}$ is said to satisfy the rank-null-space property (rNSP) of order r if for all matrices $X \in \ker(\mathcal{A})$

$$\sum_{i=1}^{r} \sigma_i(X) < \sum_{i \geq r+1} \sigma_i(X),$$

with $(\sigma_i(X))_i$ the singular values of X in decreasing order.

Proposition

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear measurement map. The two following assertions are equivalent:

1. *$\mathcal{A}$ satisfies $ER(r)$,*
2. *$\mathcal{A}$ satisfies the $rNSP(r)$.*

It is still difficult to prove the rNSP property directly. We generally circumvent this by proving a stronger assumption.

Proposition

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear measurement map. It is said to verify the rank-Restricted Isometry Property (rRIP) of order r with constant δ_r when for all matrices X of rank r,

$$(1 - \delta_r)\|X\|_{S_2}^2 \leq \|\mathcal{A}(X)\|_{\ell_m^2}^2 \leq (1 + \delta_r)\|X\|_{S_2}^2$$

Proposition

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear measurement map. There exists $c_0, c_1 > 0$ universal constant such that if $\mathcal{A}$ satisfies rRIP$(c_0 r)$ for $\delta_r < c_1$ then rNSP(r) is also satisfied.

Gaussian measurements

In the matrix case, $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ such that

$$\mathcal{A}(X) = (\langle A_i, X \rangle)_{1 \leq i \leq m}$$

is said to be a linear operator of Gaussian measurements if for all $i = 1, \ldots, m$, $A_i = (g_{k\ell})_{1 \leq k \leq d_1, 1 \leq \ell \leq d_2}$ is a Gaussian matrix, i.e. the $(g_{k\ell})$'s are i.i.d. Gaussian random variable $\mathcal{N}(0,1)$. The (A_i)'s are independent.

Theorem

Let $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear operator of Gaussian measurements and let $0 < \delta < 1$. There exist c_1, c_2, c_3 universal constants depending on δ such that with probability $1 - c_0 \exp(-c_1 m)$, $\mathcal{A}$ satisfies rRIP(r) if

$$m \geq C_2 r (d_1 + d_2).$$

What about the complexity?

Again, using a complexity argument, one can show that the necessary number m of measurements should be

$$m \gtrsim r(d_1 + d_2).$$

Summary

Norm reminder

Let H be a Hilbert space endowed with its scalar product $\langle \cdot, \cdot \rangle$. Let $E \subset H$ be a vector subspace of H and $\| \cdot \|$ be a norm on E.

- the dual norm $\| \cdot \|_*$ of $\| \cdot \|$ is defined as follows: for all $x \in E$,

$$\|x\|_* = \sup_{\substack{y \in E \\ \|y\| \leq 1}} \langle x, y \rangle.$$

- Then by definition of the dual norm, for any $x, y \in E$,

$$\langle x, y \rangle \leq \|x\| \|y\|_*.$$

- the balls and spheres associated to $\| \cdot \|$, $\| \cdot \|_*$, $\| \cdot \|_2$ (with $\| \cdot \|_2^2 = \langle \cdot, \cdot \rangle$) are noted as follows

$$\begin{aligned} B &= \{x \in E, \|x\| \leq 1\} & \qquad \mathbb{S} &= \{x \in E, \|x\| = 1\} \\ B_* &= \{x \in E, \|x\|_* \leq 1\} & \qquad \mathbb{S}_* &= \{x \in E, \|x\|_* = 1\} \\ B_2 &= \{x \in E, \|x\|_2 \leq 1\} & \qquad \mathbb{S}_2 &= \{x \in E, \|x\|_2 = 1\} \end{aligned}$$

Subdifferential of a norm

Proposition

The subdifferential of $\|\cdot\|$ at x is

$$\begin{aligned}\partial_{\|\cdot\|}(x) &= \{\eta \in E, \|x+h\| \geq \|x\| + \langle \eta, h\rangle, \forall h \in E\},\\ &= \{g \in \mathbb{S}_*, \langle g, x\rangle = \|x\|\}.\end{aligned}$$

Exercise: Show it!

Proof

If $g \in \partial_{\|\cdot\|}(x)$, then

$$
\begin{aligned}
\forall h \in E, \quad \|x+h\| - \|x\| \geq \langle g, h\rangle \\
\Rightarrow \forall h \in E, \quad \|h\| \geq \langle g, h\rangle \\
\Rightarrow \forall h \in E, \quad 1 \geq \left\langle g, \frac{h}{\|h\|} \right\rangle \\
\Rightarrow 1 \geq \sup_h \left\langle g, \frac{h}{\|h\|} \right\rangle \\
\Rightarrow \|g\|_* \leq 1.
\end{aligned}
$$

For $h = -x$, one gets $\|x\| \leq \langle g, x\rangle \Rightarrow \|g\|_* \geq 1$. Thus, $\|g\|_* = 1$. One has

$$
\begin{cases} & \|x\| \leq \langle g, x\rangle \\ \|g\|_* = 1 \Rightarrow & \|x\| \geq \langle g, x\rangle \end{cases} \Rightarrow \|x\| = \langle g, x\rangle.
$$

Proof

On the converse, let $x \in E$. If $g \in \mathbb{S}_*$ is such that $\|x\| = \langle g, x \rangle$, then for all $h \in E$,

$$\|x+h\| - \|x\| = \|x+h\|\|g\|_* - \langle g, x \rangle \geq \langle x+h, g \rangle - \langle g, x \rangle = \langle g, h \rangle.$$

Therefore $g \in \partial_{\|\cdot\|}(x)$.

The ℓ^1-norm

Rewrite the subdifferential of $\|\cdot\|_1$ at $x \in \mathbb{R}^d$ using the dual norm characterization.

$$\begin{aligned}
\partial_{\|\cdot\|_1}(x) &= \{g \in \mathbb{S}_\infty^{d-1}, \langle g, x\rangle = \|x\|_1\} \\
&= \{g \in \mathbb{S}_\infty^{d-1}, g_i = \mathrm{sign}(x_i)\} \\
&= \{g \in \mathbb{R}^d : g_i = \mathrm{sign}(x_i) \quad \text{for} \quad i \in \mathrm{supp}(x), \\
&\qquad\qquad g_i \in [-1, 1] \quad \text{otherwise}\}
\end{aligned}$$

- Therefore for x an s-sparse vector, $\partial\|\cdot\|_1(x)$ is isomorphic to a sphere $\mathbb{S}_\infty^{d-s-1}$ in dimension $\mathbb{R}^{d-s}$.
- In particular, the sparser is x, the bigger is $\partial_{\|\cdot\|_1}(x)$.
- Extreme cases are

$$\partial\|\cdot\|_1(0_{\mathbb{R}^d}) = B_\infty^{d-1} \qquad \partial\|\cdot\|_1(e_i) \subset \mathbb{S}_\infty^{d-1} \qquad \partial\|\cdot\|_1(\mathbb{1}_d) = \{\mathbb{1}_d\},$$

with $\mathbb{1}_n$ the vector of $\mathbb{R}^d$ with all components equal to 1.

Nuclear norm subdifferential

Theorem

Let $X \in \mathbb{R}^{d_1 \times d_2}$. Let $X = UDV^*$ be a singular value decomposition of X, with $D = diag(\sigma_X)$ where $\sigma_X = (\sigma_1(X), \dots \sigma_{d_1 \wedge d_2}(X))$ is the spectrum of X such that $\sigma_1(X) \geq \dots \geq \sigma_{d_1 \wedge d_2}(X) \geq 0$. Suppose X to be of rank r. Consider the following block decomposition

$$U = \begin{bmatrix} U^{(1)} & U^{(2)} \end{bmatrix} \in \mathcal{O}(d_1) \qquad D = \begin{pmatrix} diag(\sigma_1, \dots, \sigma_r) & 0_{r, d_2 - r} \\ 0_{d_1 - r, r} & 0_{d_1 - r, d_2 - r} \end{pmatrix}$$

$$V = \begin{bmatrix} V^{(1)} & V^{(2)} \end{bmatrix} \in \mathcal{O}(d_2),$$

with $U^{(1)} \in \mathbb{R}^{d_1 \times r}$ and $V^{(1)} \in \mathbb{R}^{d_2 \times r}$. The subdifferential of $\| \cdot \|_{S_1}$ at X is

$$\partial \| \cdot \|_{S_1}(X) = \{ U^{(1)} (V^{(1)})^* + U^{(2)} W (V^{(2)})^* : W \in \mathbb{R}^{(d_1 - r) \times (d_2 - r)}, \\ \|W\|_{S^\infty} \leq 1 \}.$$

Analogy with $\partial\|\cdot\|_1$ I

One can say that the term $\{U^{(1)}(V^{(1)})^*$ plays the role of the "sign" of X whereas the term $U^{(2)}W(V^{(2)})^*$ is the term (which determines how big is the subdifferential) corresponding to matrices with a different "support": for all $W \in \mathbb{R}^{(d_1-r)\times(d_2-r)}$,

$$\left\langle X, U^{(2)}W(V^{(2)})^* \right\rangle = 0.$$

One can also see the analogy by rewriting the matrix X in the $(u_i v_j^*)_{ij}$ basis (associated to some SVD of X)

$$X = UDV^* = \sum_i \sigma_i u_i v_i^*.$$

So by decomposing X in the $(u_i v_j^*)_{ij}$ basis which is an orthogonal family of rank-1 matrices, one can see that

$$\langle X, u_i v_j^* \rangle = \begin{cases} 0 & \text{if} \quad i \neq j \\ \sigma_i & \text{if} \quad i = j. \end{cases}$$

Analogy with $\partial\|\cdot\|_1$ II

So, one can write any subgradient of $\|\cdot\|_{S_1}$ at X such that

$$\sum_{i=1}^{r} u_i v_i^* + U^{(2)} W (V^{(2)})^*$$

for $W \in \mathbb{R}^{(d_1-r)\times(d_2-r)}$ such that $\|W\|_{S^\infty} \leq 1$. Here the term $U^{(2)} W (V^{(2)})^*$ can be seen as having a different support of X.

Sparse recovery vs. low-rank minimization

Object	Vectors	Matrices
Structure	Sparsity (s)	Low-rank (r)
Hilbert space norm	Euclidean	Frobenius
Sparsity-inducing norm	ℓ_1-norm	nuclear norm
Dual norm	ℓ_∞-norm	operator norm $\\|\cdot\\|_{2\to 2}$
Norm additivity	disjoint supports	orthogonal row and column spaces
Convex optimization	Linear programming	SDP

On the descent cone

Recall that x is a minimizer if the descent cone $\mathcal{D}$ at x trivially intersect the null space of A, i.e.

$$\mathcal{D} \cap \ker(A) = \{0\},$$

where $\mathcal{D} = \mathcal{D}_{\|\cdot\|}(x)$ is the descent cone. If x is not optimal for $\|\cdot\|$ in the sense that $x \neq 0$, then the descent cone coincides with the polar cone of the cone spanned by the subgradients.

Exercise

For the $\ell^1 -$ norm, and for the points in the following Figure ,

1. *draw the descent cones at the following points,*
2. *draw the cone spanned by their subgradients for the ℓ_1-norm,*
3. *conclude.*

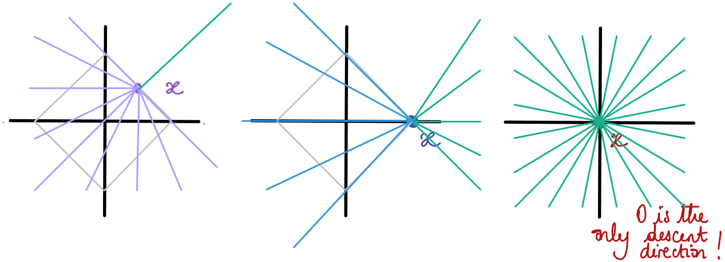

Figure: Comparison with the cone of the subgradients (in green): except in the case where x is optimal for the $\|\cdot\|_1$, i.e. $x \neq 0$, the descent cone coincides with the polar cone to the cone spanned by the subgradients. (if $x = 0$ its descent cone is $\{0\}$ which does not admit a polar cone).

Recall that for a cone $C \subseteq X$, the polar cone is given by

$$C^\circ = \{y \in X^*, \langle y, x\rangle \leq 0 \quad \forall x \in X\}.$$

Wrapping-up

If $x \neq 0$,

$$\mathcal{D}(x, \|\cdot\|) = \left(\text{cone}\,\{\eta, \eta \in \partial\|\cdot\|(x)\}\right)^{\circ}.$$

Conclusion

The sparser the point of interest,

- the bigger the cone spanned by its subgradients,
- the smaller its polar cone,
- i.e. the smaller the descent cone at this point is,
- the easier is satisfied the condition

$$\mathcal{D} \cap \ker(A) = \{0\}.$$

Can someone draw the nuclear ball?

Drawing the nuclear ball I

Consider a symmetric matrix:

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}.$$

where a, b, c are real numbers. Compute the singular values $\sigma_{1,2}$ of A as positive square roots of the eigenvalues of $A^T A = A^2$:

$$\sigma_{1,2} = \frac{1}{\sqrt{2}}\sqrt{a^2 + 2b^2 + c^2 \pm |a+c|\sqrt{(a-c)^2 + 4b^2}}.$$

A is in the nuclear ball if $\sigma_1 + \sigma_2 < 1$. Square the previous expression

$$a^2 + 2b^2 + c^2 + \sqrt{(a^2 + 2b^2 + c^2)^2 - (a+c)^2((a-c)^2 + 4b^2)} = 1$$

and so

$$a^2 + 2b^2 + c^2 + 2|b^2 - ac| = 1.$$

Drawing the nuclear ball II

Thus if $ac \geq b^2$, then
$$(a+c)^2 = 1,$$
(which gives the planar caps) if $ac < b^2$, then
$$(a-c)^2 + 4b^2 = 1,$$
which gives the cylinder.
If one considers matrices such that
$$\begin{cases} (a+c)^2 & = 1 \\ (a-c)^2 + 4b^2 & = 1 \end{cases}$$
one can show that $b^2 = ac$, which leads to a matrix of rank one, since
$$\begin{pmatrix} x \\ z \end{pmatrix} \begin{pmatrix} x & z \end{pmatrix} = \begin{pmatrix} x^2 & xz \\ xz & z^2 \end{pmatrix}.$$

Drawing the nuclear ball III

Therefore matrices such that

$$\begin{cases} (a+c)^2 & = 1 \\ (a-c)^2 + 4b^2 & = 1 \end{cases}$$

are matrices of the extremal ellipses of the unit nuclear ball, which corresponds to rank-1 matrices.

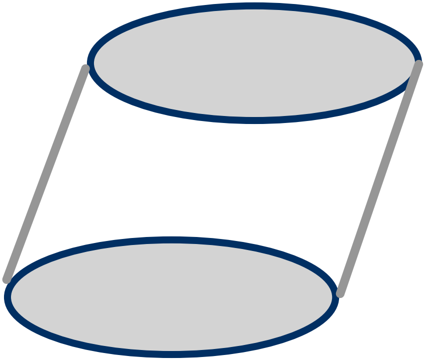

Figure: Unit ball asocciated to the nuclear norm: the extremal ellipsoids (in blue) corresponds to rank-1 matrices, with spectral norm bounded by 1.

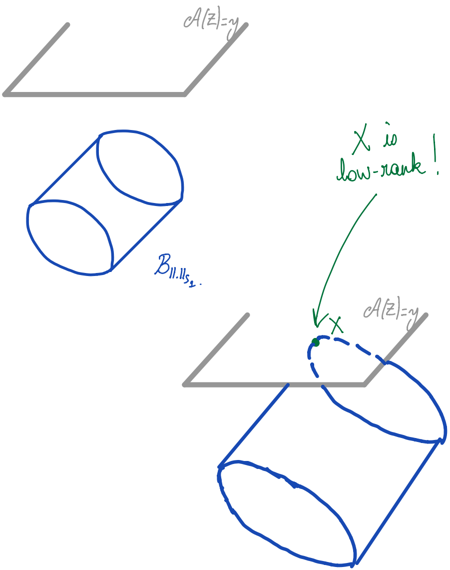

Figure: Nuclear norm minimization: the solution will be low-rank.

Summary

1. Preliminary
   - SVD reminder
   - Scalar product and norms
   - Sets
2. Convexification
3. Exact reconstruction for (NNM)
4. (NSP,RIP)-based reconstruction guarantee
5. Non-differentiability of the objective function
   - Subdifferential of a norm
   - Subdifferential of the nuclear norm
   - Complexity of the subdifferential
   - Drawing the nuclear ball
6. Matrix completion
   - Setting
   - Non-uniform recovery results
     - Coherence
     - Nonuniform results
   - What about nonuniform sampling?
   - Algorithm: proximal method for matrix completion

Setting of matrix completion

- Goal: to fill in missing entries of a low-rank matrix.
- The measurement map $\mathcal{A}$ samples the entries $\mathcal{A}(X)_\ell = X_{j,k}$ for some indices (j,k) depending on ℓ. Let's say that we can see the entries of X that are indexed by $\Omega \subset [|d_1|] \times [|d_2|]$. So, the entries of X are revealed uniformly at random so that the sampling "vectors" (A_ℓ) can be written as $(e_i e_j^*)_{(i,j)\in\Omega}$, so

$$\forall (i,j) \in \Omega, \quad X_{i,j} = \langle e_i e_j^*, X \rangle,$$

or using a projection on a mask Ω:

$$P_\Omega : \begin{cases} \mathbb{R}^{d_1\times d_2} & \to \mathbb{R}^{d_1\times d_2} \\ X & \to P_\Omega(X) \end{cases} \quad \text{with} \quad (P_\Omega(X))_{ij} = \begin{cases} X_{ij} & \text{if} \quad (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

- Set the measurement vector $\qquad y = P_\Omega X \in \mathbb{R}^m.$

Application

- Recommendation system: customer taste prediction
- $\rightsquigarrow$ Netflix prize

Incoherence for matrix completion

Need a notion of incoherence between the sampling vectors (of the form $e_i e_j^*$) and the matrix X.

Definition (Incoherence for matrix completion)

For this setting, define the incoherence parameter $\mu(X)$ as the smallest number such that

$$\max_{1\leq i\leq d_1} \left(\frac{d_1}{r}\right) \cdot \|P_{\mathrm{col}(X)} e_i\|_2^2 \leq \mu(X)$$
$$\max_{1\leq j\leq d_2} \left(\frac{d_2}{r}\right) \cdot \|P_{\mathrm{row}(X)} e_i\|_2^2 \leq \mu(X),$$

where $r = \mathrm{rank}(X)$, $P_{\mathrm{col}(X)}$ (resp. $P_{\mathrm{row}(X)}$) is the projection onto the column (resp. row) space of X.

The coherence parameter measures the overlap or correlation between the column/row space of the matrix and the coordinate axes.

Incoherence in words

Since

$$\sum_i \|P_{\mathrm{col}(X)} e_i\|_2^2 = \mathrm{Tr}(P_{\mathrm{col}(X)}) = r,$$

one can conclude that $\mu(X) \geq 1$.
Conversely $\mu(X) \leq \max(d_1, d_2)/r$.

- A matrix with low coherence has column and row spaces away from the coordinate axes as in the case where they assume a uniform random orientation. If the column space of X has uniform orientation then for each i

$$\frac{d_1}{r}\mathbb{E}\|P_{\mathrm{col}(X)} e_i\|_2^2 = 1.$$

- Conversely, a matrix with high coherence may have a column (or a row space) well aligned, with a coordinate axis. As should become intuitive, we can only hope to recover "incoherent" matrices; i.e. matrices with relatively low-coherence parameter values.

Non-uniform result

Theorem

Let X be a fixed but otherwise arbitrary matrix of size $d_1 \times d_2$ and of rank r. Let y be the set of revealed entries of X at randomly selected locations. Then with probability $1 - n^{-10}$, X is the unique minimizer of (NNM)*, provided that the number m of samples obeys*

$$m \geq c_0 \cdot \mu(X) \cdot r(d_1 + d_2) \cdot \log^2(d_1 + d_2),$$

for some positive universal constant c_0.

Coherence and movie rating

One can ask whether matrix completion is possible from more general random equations, where the sampling matrices may not have rank one, and are still i.i.d. samples from some fixed distribution F.
To give a concrete example, suppose we have an orthobasis of matrices $\mathcal{B} = \{B_j\}_{1 \leq j \leq d_1 d_2}$, and that we select elements from this family uniformly at random. Then [Gross,2011] shows that if

$$\max_{B \in \mathcal{B}} \left(\frac{d_1}{r}\right) \cdot \|P_{\mathrm{col}(X)} B\|_F^2 \leq \mu(X)$$
$$\max_{B \in \mathcal{B}} \left(\frac{d_2}{r}\right) \cdot \|P_{\mathrm{row}(X)} B\|_F^2 \leq \mu(X),$$

holds along with another technical condition, the nonuniform recovery result presented above holds.
Note that in the previous example where $B = e_i e_j^*$, $\|P_{\mathrm{col}(X)} B\|_F^2 = \|P_{\mathrm{col}(X)} e_i\|_2^2$ so that we are really dealing with the same notion of coherence.

Regularized problem

We want to use proximal methods to solve the matrix completion problem. One observes the partial measurements $P_\Omega(X)$, and we suppose that X is low-rank. One can consider the two following problems:

1. Find (constrained version with the rank)

$$X^\sharp \in \underset{\mathrm{rank}(Z)\leq r}{\arg\min} \; \|P_\Omega(Z) - P_\Omega(X)\|_{S_2}. \tag{1}$$

2. Find (convex regularized/Lagrangian version)

$$X^\sharp \in \arg\min_Z \frac{1}{2}\|P_\Omega(Z) - P_\Omega(X)\|_{S_2}^2 + \lambda\|Z\|_{S_1} \tag{2}$$

To implement these procedures, we are going to consider "hard-thresholding" and "soft-thresholding"-like approaches.

Prox of the nuclear norm

Proposition

Let $\gamma > 0$. The proximal operator of $\gamma\|\cdot\|_{S_1}$ is the spectral soft-thresholding operator:

$$\mathrm{prox}_{\gamma\|\cdot\|_{S_1}}(X) = S_\gamma(X) = UD_\gamma V^T,$$

with $X = UDV^T$ the SVD of X and

$$D_\gamma = diag((\sigma_1 - \gamma)_+, (\sigma_2 - \gamma)_+, \ldots)$$

Proof: Blackboard time.

Proximal gradient descent for NNM

Once the proximal operator computed, one can construct a proximal gradient descent to solve (2).

Proximal gradient descent

1. $F(X) = (1/2)\left\|P_\Omega(B) - P_\Omega(A)\right\|_{S_2}^2$,
   $\nabla F(X) = P_\Omega^\top(P_\Omega(B) - P_\Omega(X))$
2. $X_{k+1} = \mathrm{prox}_{\gamma_k\|\cdot\|_{S_1}}\left(X_{k+1} - \gamma_k(P_\Omega^\top(P_\Omega(B) - P_\Omega(X))\right)$ where $(\gamma_k)_k$ is a size step family.

Hard-thresholding algorithm

One can also solve (1), one has to replace the soft-thresholding operator by the hard-thresholding operator:

$$\mathop{\arg\min}_{B\in\Sigma_r} \|X - B\|_{S_2} = \{UD_r V^\top\}$$

where $X = UDV^\top$ and $D_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$. The algorithm then follows

Algorithm

$$X^{k+1} = H_\gamma\big(X^k - \gamma_k P_\Omega^\top(P_\Omega(B) - P_\Omega(X))\big)$$

where H_μ is the hard-thresholding operator:

$$H_\mu(X) = UD_\mu V^\top \text{ with } D_\mu = \mathrm{diag}(\sigma_1\delta_{\sigma_1\geq\mu}, \sigma_2\delta_{\sigma_2\geq\mu}, \ldots).$$

Summary

Phase retrieval I

Application with phase retrieval appearing

- speech analysis
- astronomical imaging
- photographic plates, CCDs and other light detectors
- measure the intensity of an electromagnetic wave as opposed to measuring its phase
- example of X-ray crystallography: a collimated beam of X-rays strikes a crystal; these rays then get diffracted by the crystal or sample and the intensity of the diffraction pattern is recorded.

Phase retrieval II

Measurements

If $x(t_1, t_2)$ is a discrete two-dimensional object of interest, data is of the form

$$y(\omega_1, \omega_2) = \left| \sum_{t_1, t_2}^{d} x(t_1, t_2) e^{2i\pi(t_1\omega_1 + t_2\omega_2)} \right|^2, \quad (\omega_1, \omega_2) \in \Omega,$$

with Ω a sampled set of frequencies in $[0, 1]^2$.

Question

- how one can invert the Fourier transform from phaseless measurements?

Phase retrieval III

The abstract formulation of the phase retrieval problem asks us to solve a system of quadratic equations,

$$y_k = |\langle a_i, x \rangle|^2, \quad k = 1, \ldots, m \tag{3}$$

in which x is an d-dimensional complex or real-valued object. This is quite different from the underdetermined linear systems considered thus far. In passing, solving quadratic equations is known to be notoriously difficult (NP-hard).

Cast the phase retrieval pb as a matrix completion one

To see this, introduce the $d \times d$ positive semidefinite Hermitian matrix variable $X \in \mathcal{S}^d$ equal to xx^*, and observe that

$$|\langle a_i, x \rangle|^2 = \mathrm{Tr}(a_k a_k^* x x^*) = \mathrm{Tr}(A_k X) \qquad \text{with} \quad A_k = a_k a_k^*. \tag{4}$$

Phase retrieval IV

By lifting the problem into higher dimensions, we have turned quadratic equations into linear ones!

Suppose that (3) has a solution x_0. Then there obviously is a rank-one solution to the linear equations in (4), namely, $X_0 = x_0 x_0^*$. Thus the phase retrieval problem is equivalent to finding a rank-one matrix from linear equations of the form $y_k = \mathrm{Tr}(a_k a_k^* X)$. This is a rank-one matrix completion problem! Since the nuclear norm of a positive definite matrix is equal to the trace, the natural convex relaxation:

PhaseLift

$$\min \mathrm{Tr}(X) \qquad \text{s.t.} \qquad X \succcurlyeq 0, \quad \mathrm{Tr}(a_k a_k^* X) = y_k, k = 1, \ldots, m. \tag{5}$$

Phase retrieval V

Theorem

Suppose the (a_k)'s are independent random vectors uniformly distributed on the sphere - equivalently, independent complex-valued Gaussian vectors- and let $\mathcal{A} : \mathbb{C}^{d\times d} \to \mathbb{R}^m$ be the linear map $\mathcal{A}(X) = (\mathrm{Tr}(a_k a_k^* X))_{1\leq k\leq m}$. Assume that

$$m \geq c_0 d,$$

where c_0 is a sufficiently large constant. Then the following holds with probability at least $1 - O(e^{-\gamma m})$: for all x_0 in $\mathbb{C}^d$, the feasibility problem

$$\{X : X \succcurlyeq 0 \qquad \text{and} \qquad \mathcal{A}(X) = \mathcal{A}(x_0 x_0^*)\}$$

has a unique point, namely, $x_0 x_0^*$. Thus, with the same probability, PhaseLift recovers any signal $x_0 \in \mathbb{C}^d$ up to a global sign factor.

Phase retrieval VI

- This theorem states that a convenient convex program (SDP) can recover any n-dimensional complex vector from on the order of n randomized quadratic equations.
- In short, convex programming techniques and matrix completion ideas can be brought to bear, with great efficiency, on highly non-convex quadratic problems

Robust PCA I

PCA

Suppose we have a family of n points belonging to a high-dimensional space of dimension d, which we regard as the columns of a $d \times n$ matrix M. Many data analysis procedures begin by reducing the dimensionality by projecting each data point onto a lower dimensional subspace. Principal component analysis (PCA) achieves this by finding the matrix X of rank k, which is closest to M in the sense that it solves:

$$\min \|M - X\| \qquad \text{s.t.} \qquad \operatorname{rank}(X) \leq k.$$

where $\|\cdot\|$ is either the Frobenius or the usual spectral norm.

Robust PCA II

The solution is given by truncating the singular value decomposition as to retain the k largest singular values. When our data points are well clustered along a lower dimensional plane, this technique is very effective.

- In many real applications, many entries of the data matrix are typically either unreliable or missing:
  - entries may have been entered incorrectly,
  - sensors may have failed,
  - occlusions in image data may have occurred, and so on.
- The problem is that PCA is very sensitive to outliers and few errors can throw the estimate of the underlying low-dimensional structure completely off. Researchers have long been preoccupied with making PCA robust.

Robust PCA III

Imagine we are given a $d \times n$ data matrix

$$M = L_0 + S_0$$

where L_0 has low rank and S_0 is sparse.
We observe M but L_0 and S_0 are hidden. The connection with our problem is that we have a low-rank matrix that has been corrupted in possibly lots of places but we have no idea about which entries have been tampered with. Can we recover the low-rank structure? The idea is to de-mix the low-rank and the sparse components by solving:

$$\min \|L\|_{S^1} + \lambda \|S\|_{\ell^1} \qquad \text{s.t.} \qquad M = L + S \tag{6}$$

Robust PCA IV

Theorem

Assume without loss of generality that $n \geq d$, and let L_0 be an arbitrary $n \times d$ matrix with coherence $\mu(L_0)$ as defined before. Suppose that the support set of S_0 is uniformly distributed among all sets of cardinality m. Then with probability at least $1 - O(n^{-10})$ (over the choice of support of S_0), (L_0, S_0) is the unique solution to (6) with $\lambda = 1/\sqrt{n}$ provided that

$$\operatorname{rank}(L_0) \leq c_0 \cdot d \cdot \mu(L_0)^{-1} (\log n)^2 \qquad \text{and} \qquad m \leq c_1 \cdot n \cdot d,$$

with c_0 and c_0' numerical constants.

Hence, if a positive fraction of the entries from an incoherent matrix of rank at most a constant times $d/\log^2(n)$ are corrupted, the convex program (6) will detect those alterations and correct them automatically.