High-Rank Matrix Completion and Subspace Clustering with Missing Data

Authors: Brian Eriksson, Laura Balzano and Robert Nowak

Presentation by Wang Yuxiang

About the authors

Brian Eriksson
Postdoctoral Research Fellow
Boston University (Computer Science)
University of Wisconsin (Electrical Engineering)

<-**Laura Balzano**
Phd student ->

Robert Nowak->
Professor, ECE,
U Wisc.Madison

The recent works of Nowak's group

- GROUSE, 2010
- GRASTA, 2011
- High dimensional matched subspace detection when data are missing, 2010
- K-subspace with missing data, 2011

Preliminary results used in this paper

- A simpler approach of matrix completion, 2011, Ben Recht
  - Lemma 6 in this paper.
- High dimensional matched subspace detection when data are missing, 2010
  - Used in Lemma 8 in this paper.

Outline of presentation

- Problem definition/motivations
- Stages of the algorithm
- Key results and discussion
  - Assumptions
  - Theorem
  - Discussion
- Simulation and real data experiment
- Stages of the proof

Before presenting the paper

- A LOT of terminologies, parameters.

- Feel free to stop me any time to get back on track.

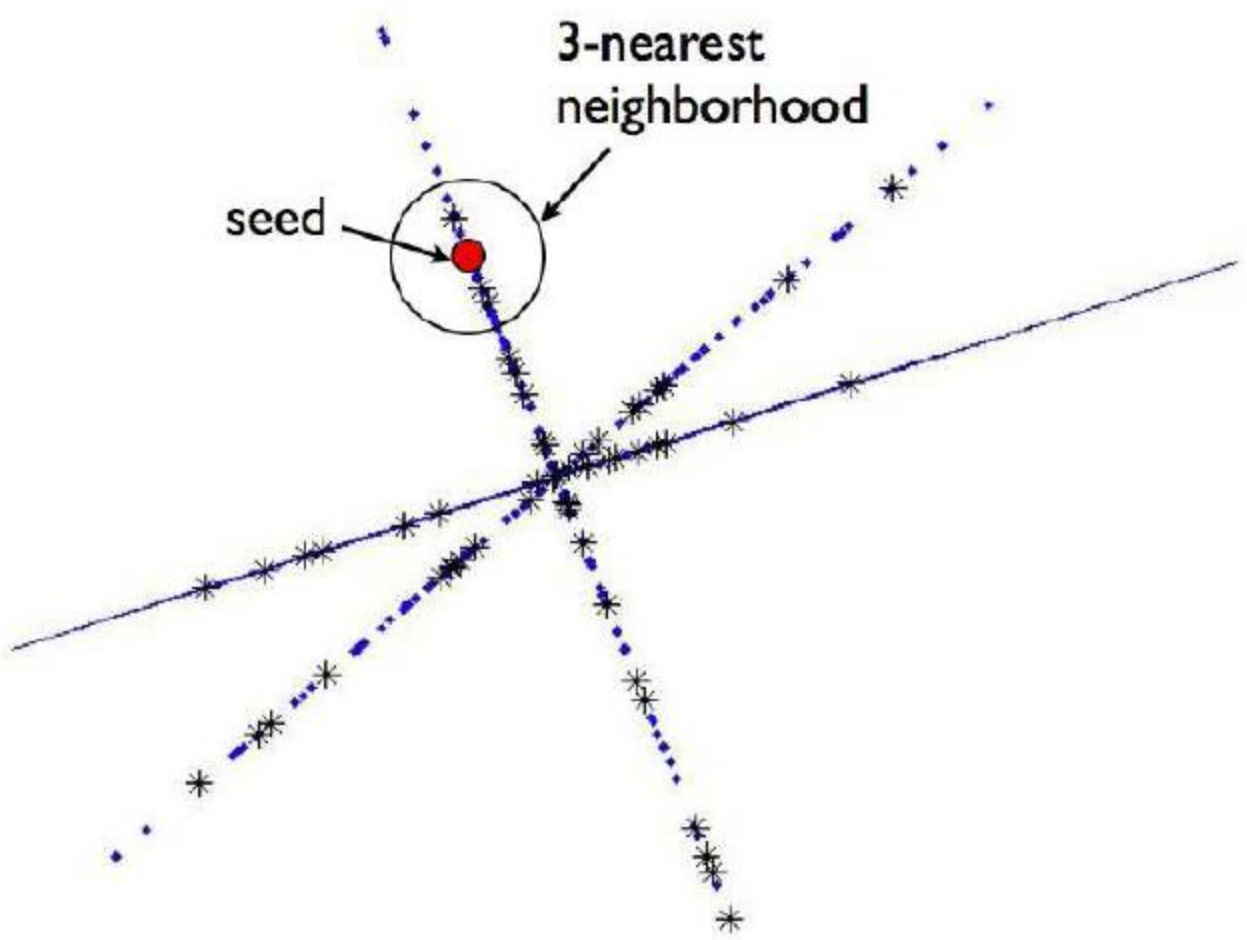
High rank matrix completion: definition

- Given n by N matrix X, columns of X lie in the union of k subspaces in R^n. If only a subset of indices Ω are observed from X, under what conditions can original X be fully recovered and how?
- Key observations:
  - In general, this is not possible (need assumptions)
  - Much more difficult than low rank MC. (chicken-egg)
  - Potentially, X can be full rank.

Motivation

- In computer vision:
  - Motion segmentation with partial feature track
  - Articulated/deformable object motion
  - Face recognition, object recognition
- In other fields:
  - Collaborative filtering (Netflix)
  - Network Topology inference (motivating application of this paper)
  - Anything else that is well approximated by hybrid linear model

An illustration of hybrid linear data



A sketch of the algorithm

- Step 1: Local Neighborhoods
  - Find seeds and establish neighborhood
- Step 2: Local Subspaces
  - Do matrix completion for each seed/neighborhood
- Step 3: Subspace Refinement
- Step 4: Full Matrix Completion
  - Column membership classification
  - Complete each column from subspace

Summary of results

- The paper demonstrated that exact completion of each column is possible with high probability.
- What conditions?

  If sample rate $$p_0 \geq C \frac{r}{n} \log^2(n)$$

  # of columns $$N \gg kn$$

  and the three "mild" assumptions on the structure of the matrix is satisfied.

  Moreover if we change log(n) to log(N), then the result extends to exact completion of the FULL matrix.

Assumptions

- What conditions?

A1. k= o(n^d), each subspace is at most rank r, each column has l_2-norm <= 1.

(used in Lemma 7)

A2. μ_0 incoherent subspace (vis a vis standard basis)

μ_1 incoherent columns AND difference of any two columns (no spiky entries)

(Used in Lemma 3)

A side remark on Incoherence property

- Coherence of a subspace S:

$$\mu(\mathcal{S}) := \frac{n}{r} \max_j \|P_{\mathcal{S}} e_j\|_2^2 \qquad 1 \leq \mu(\mathcal{S}) \leq n/r$$

- Coherence of a vector/column x:

$$\mu(x) = \frac{n\|x\|_\infty^2}{\|x\|_2^2}$$

- This paper requires columnwise μ_1 incoherence and the difference of columns to be μ_1 incoherent too, which is rather restrictive.

A side remark on Incoherence property

- Assumptions in Recht's Simpler MC:
  - Subspace incoherence μ_0 (Same as here)
  - Row/column cross subspace incoherence μ_1 defined as:

    The matrix $\boldsymbol{UV}^*$ *has a maximum entry bounded by* $\mu_1\sqrt{r/(n_1 n_2)}$ *in absolute value*.

- This is different from the μ_1 here! The authors blindly used it anyway.
- The definition is more restrictive, may implies the condition of Recht's (μ_1)(?)

Assumptions

- What conditions?

  A3. Matrix X is sufficiently random
  - No columns lies in intersection of two subspace (hence no ambiguity). (used in Lemma 8)
  - Any r_i columns of Subspace S_i spans the subspace.(Used in subspace refinement step)
  - Any two columns in different subspaces are at least ε_0 away. (used in the arguments following Lemma 3)
  - Random select j<{1,2,…,N}, min(Prob(j \in $S_{i,\varepsilon 0}$)) >= ν_0/k, (used in Lemma 2)
  - for any column x belonging to any subspace S_i, random select a column j, then Prob(||X_j - x|| <= ε_0)>= ν_0 ε_0^r/k (used in Lemma 4)

Summarizing the parameters

- n, N, k, r
  - Size of matrix, number of subspace, max rank
- μ_0, μ_1
  - Coherence of subspace and that of each column and column difference.
- ε_0, ν_0
  - Min separation (in Euclidean distance), skewness of subspace sampling.
- η_0, t_0, s_0, l_0, p_0
  - Min seed sampling, min overlap with neighbors, number of seeds to be chosen, random # of columns to guarantee, rate of random sampling

Main Theorem (Thm2.1)

- Define the following quantities:

$$\delta_0 := n^{2-2\beta^{1/2}} \log n \,, \text{ for some } \beta > 1 \,,$$

$$s_0 := \left\lceil \frac{k(\log k + \log 1/\delta_0)}{(1-e^{-4})\nu_0} \right\rceil \,,$$

$$\ell_0 := \left\lceil \max \left\{ \frac{2k}{\nu_0 (\frac{\epsilon_0}{\sqrt{3}})^r} \,, \frac{8k \log(s_0/\delta_0)}{n\nu_0 (\frac{\epsilon_0}{\sqrt{3}})^r} \right\} \right\rceil \,.$$

Main Theorem (Thm2.1)

Let X be an n by N Matrix satisfying A1-A3, given iid entrywise observation under sample rate p_0,

If $$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \frac{r\,\log^2(n)}{n}$$

and $$N \geq \ell_0 n (2\delta_0^{-1} s_0 \ell_0 n)^{\mu_0^2 \log p_0^{-1}}$$

Then each column can be exactly recovered

$$\textit{with probability at least } 1 - (6 + 15 s_0)\,\delta_0.$$

Discussion on the results

- It does not require exact k to be known in prior.

- It does not require independent subspaces. Overlaps are allowed as long as the separation assumptions are satisfied.

Discussion on the results

- Is the sample rate requirement optimal?
  - Degree of freedom of the problem: O(knr)
  - Current result is $O(Nr\log^2(n))$, hence already near optimal (Note that N is dependent on k)
  - At full rank: kr>n, d.o.f. > n^2, the result is $O(Nn\log^2(n)/k)$, so it still makes sense.
- The possible improvement is on the N. Can we do high rank matrix completion on square matrix?

Discussion on the results

- Restrictive assumptions:
  - The μ_1 condition as discussed before.
  - Extremely large N (number of samples) is required. To get an inkling:

At constant fraction sample rate:

$$N=O(poly(kn/\delta_0))$$

At diminishing sample rate, say $O(\log^2(n)/n)$:

$$N = O((kn/\delta_0)^{poly(log(n))})$$

Discussion of the results

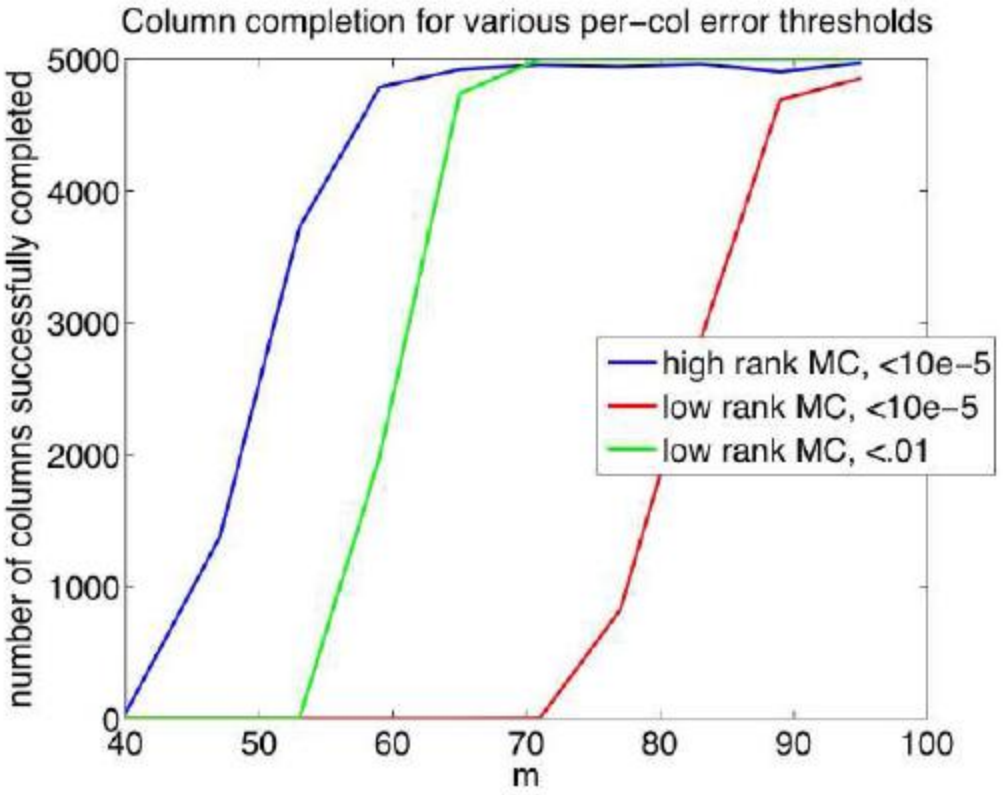
- Verification of assumptions is either NP-hard, or relies on data not available, or both.

- Some of assumptions might be redundant, e.g., the any r columns span subspace condition should hold given incoherence property.

Simulations: Compare to low-rank MC

- n= 100, N=5000, k=10, r=5
- Each r-dim subspace is generated by span of n*r gaussian random matrix, U: orthonormal basis.
- 500 samples are drawn from each subspace from a Gaussian distribution N(0,UU*).
- Number of seeds are chosen to be 3klog(k)
- Standard MC is conducted using GROUSE (!)

Simulations: Compare to low-rank MC



- Per-column samples m
- Requirement of High rank MC:

  r log(n) = 23
- Requirement of Low rank MC:

  kr log(n) = 230, since total rank is kr.

Additional Simulations: Compare to GROUSE and Nuclear Norm

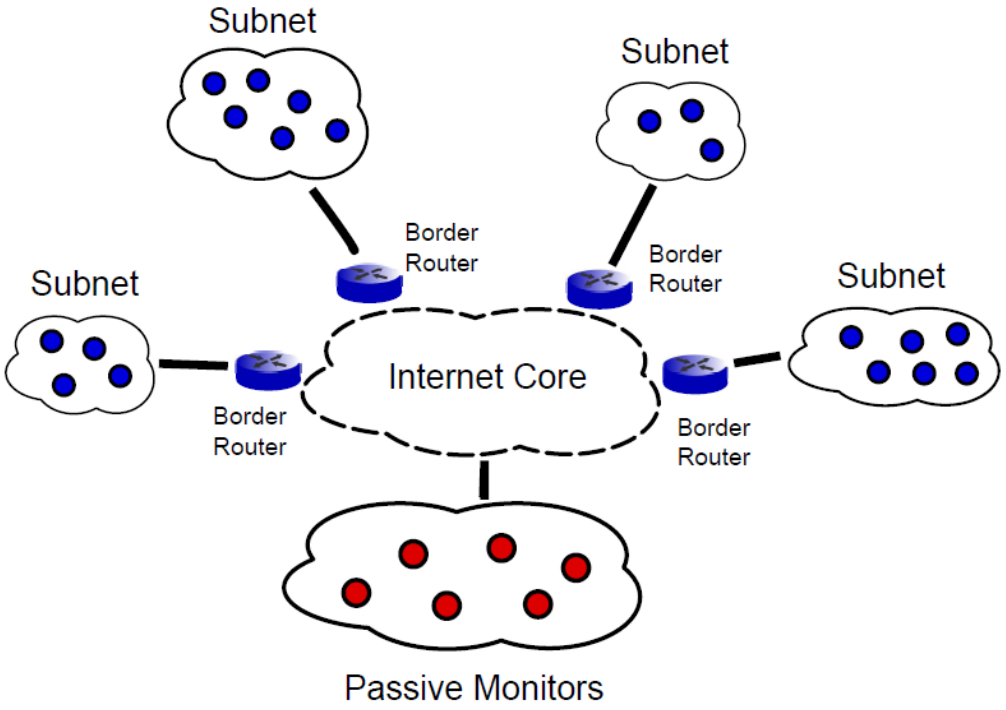
	0.5	0.6	0.7	0.8	0.9	
500	0	0	0	100.00%	100.00%	Grouse
	0	0	0	99.50%	99.50%	Nuclear
100	0	0	0	0	86.90%	Grouse
	0	0	0	99.90%	99.90%	Nuclear
50	0	0	0	0	0	Grouse
	0	0	0	99.80%	100.00%	Nuclear
10	0	0	0	0	0	Grouse
	0	0	0	0	0	Nuclear

- Horizontal axis is sample rate.
- Vertical axis is number of columns per subspace, so from top to bottom number of columns is 5000,1000, 500, 100.
- Nuclear norm based MC is performed using TFOCS(Template for First Order Conic Solver) to eliminate the possible numerical issues of APG.
- GROUSE is performed using their released code.

Network Topology Inference Experiments

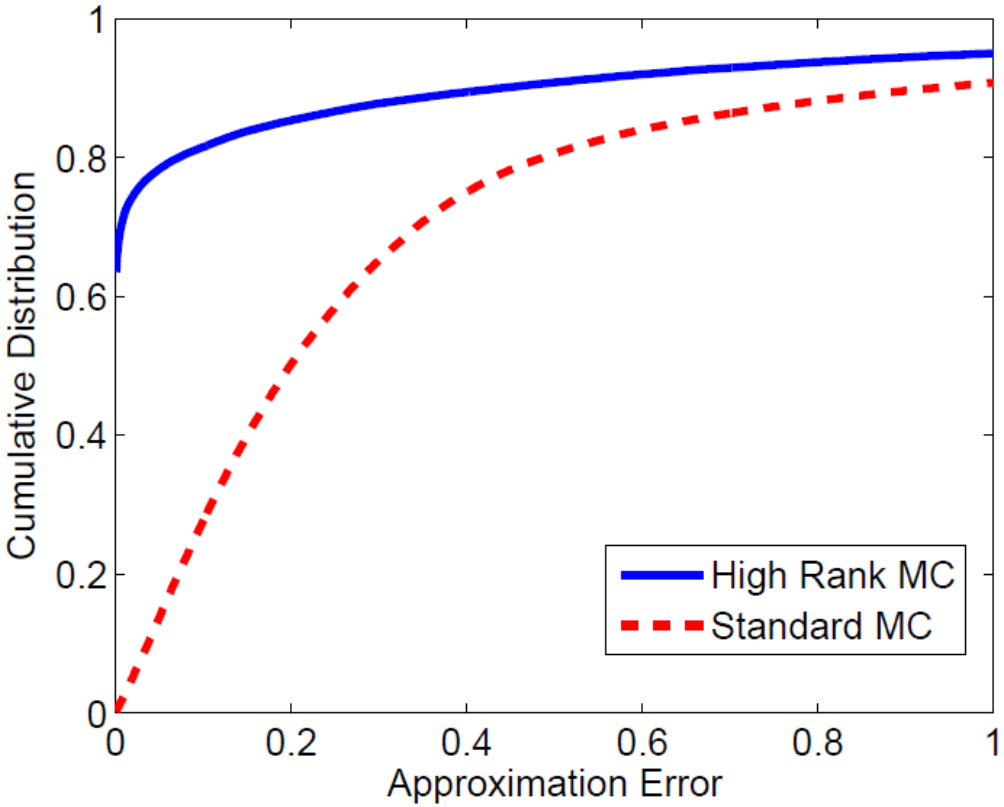
- Recover router-level connectivity without active probing. Using incomplete, passively observed measurements.
  - n is the number of monitors,
  - N is the total unique IP addresses observed,
  - the value of each entry represents hop-counts.
- This matrix should consist of a union of many rank-2 subspace. (Why?)

Network topology illustration



- All IP addresses/columns in a subnet is rank 2.
- Because any probe sent from an IP in a subnet must traverse through the same border router.
- Each hop count vector = border router's hop count vector + constant offset
- The constant offset is related to the distance from each IP to border router.

Simulation of network topology inference



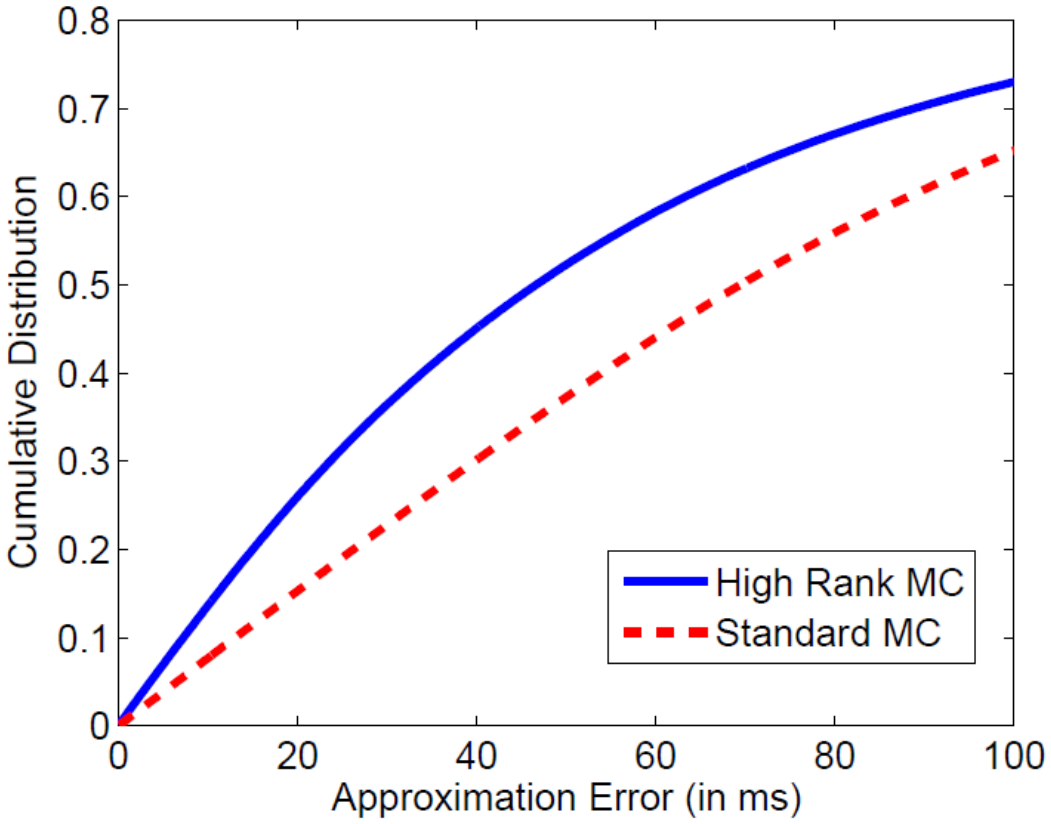
A synthetic network with:

k=12 subnets, n=75 passive monitors, N=2700 ip addresses

Each subnet has dimension r=2.

40% of the elements are observed.

Real network data experiment



- n=100
- N=22550
- k is unknown, but the parameters are estimated with an estimation of k=15
- delay time is used as an estimate of hop count.(so with noises!)
- Roughly 40% of total delay time is observed.

Stages of the algorithm/proof

- Step 1: Local Neighborhoods
- Step 2: Local Subspaces
- Step 3: Subspace Refinement
- Step 4: Full Matrix Completion

This stage by stage structure is followed by both the algorithm and proof.

Concentration inequalities

- Chernoff bound

$$P(\bar{X}_n \geq \mu + \epsilon) \leq \exp[-n\epsilon^2/2(\mu+\epsilon)] \qquad P(\bar{X}_n \leq \mu - \epsilon) \leq \exp[-n\epsilon^2/2\mu].$$

$$P(\bar{X}_n \geq (1+\delta)\mu) \leq \exp[-n\mu\frac{\delta^2}{2(1+\delta)}] \qquad P(\bar{X}_n \leq (1-\delta)\mu) \leq \exp[-n\mu\delta^2/2]$$

- Hoeffding's Inequality

$$\Pr(\overline{X} - \mathrm{E}[\overline{X}] \geq t) \leq \exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right), \quad \Pr(\overline{X} - \mathrm{E}[\overline{X}] \geq t) \leq \exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right),$$

- McDiramid's Inequality

$$\forall i, \forall x_1, \ldots, x_m, x_i' \in \mathcal{X}, \quad |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i.$$

Then for all $\epsilon > 0$,

$$\Pr\left[f - \mathbb{E}[f] \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\right).$$

Step1: Local neighborhood procedure

- Input: $n, k, \mu_0, \epsilon_0, \nu_0, \eta_0, \delta_0 > 0.$

- Parameters:

$$
\begin{aligned}
s_0 &:= \left\lceil \frac{k(\log k + \log 1/\delta_0)}{(1-e^{-4})\nu_0} \right\rceil \\
\ell_0 &:= \left\lceil \max\left\{ \frac{2k}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}, \frac{8k\log(s_0/\delta_0)}{n\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \right\} \right\rceil \\
t_0 &:= \lceil 2\mu_0^2 \log(2s_0\ell_0 n/\delta_0) \rceil
\end{aligned}
$$

Step1: Local neighborhood procedure

- Procedures:
  1. Randomly select s_0 seeds with at least η_0 samples.
  2. For each seeds, find all columns with t_0 overlaps
  3. Randomly select $l_0 n$ columns from each set
  4. From these $l_0 n$ columns, randomly select n columns with partial distance less than $\varepsilon_0 /sqrt(2),$ which forms the local neighbor sets.

Why would this procedure work?

- Lemma 2: If we randomly take sufficient number of columns, then there are at least one seed for each subspace that has sufficient observation with probability $1-\delta_0$

$$s \geq \frac{k(\log k + \log 1/\delta_0)}{(1-e^{-4})\nu_0}, \quad \eta_0 := \frac{64\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0}\, r \cdot \log^2(n)$$

- Proof requires assumption A3.

Why would this procedure work?

- Lemma 3: For any two column x1, x2, y= x1-x2, common observation set is ω, if common observations:

$$q \geq 8\mu_1^2 \log(2/\delta_0) ,$$

then with probability 1-δ_0

$$\frac{1}{2}\|y\|_2^2 \leq \frac{n}{q}\|y_\omega\|_2^2 \leq \frac{3}{2}\|y\|_2^2 .$$

Why would this procedure work?

- Implication of Lemma 3: subspace membership can be seen from partial distance!
  - We know distance between data in different subspace is at least ε_0.
  - If $|x1-x2|^2 <= \varepsilon_0^2/3$, then $n/q|y_\omega|^2 <= \varepsilon_0^2/2$

    Can be used to construct the conditions of having desired observations.
  - If $n/q|y_\omega|^2 <= \varepsilon_0^2/2$, then $|x1-x2|^2 <= \varepsilon_0^2$

    Can be used to infer subspace membership from partial observations.

Why would this procedure work?

- Lemma 4: If we randomly sample sufficient number (*ln*) of columns for each seeds,

$$\ell \geq \max\left\{ \frac{2k}{\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r}, \frac{8k\log(s/\delta_0)}{n\nu_0(\frac{\epsilon_0}{\sqrt{3}})^r} \right\},$$

  with probability 1-δ_0, there is at least n columns within $\epsilon_0/\sqrt{3}$ of all s seeds.

- So these columns are not only of the same subspace, but also manifests this info through the partially observed distance!

Why would this procedure work?

- Lemma 5: Gluing Lemma 2,3,4 together. If N being sufficiently large and $\eta_0 > t_0$ then the "Local neighborhood procedure" produces at least n columns within $\epsilon_0/\sqrt{3}$ of each seeds, and at least one seed belongs to each subspace with probability 1-3δ_0.
- The proof applies Lemma 2, Lemma 4 with δ_0 then Lemma 3 with $\delta_0/(s_0 l_0 n)$, then use union bound to get the 3δ_0.

Why would this procedure work?

- Lastly, N must be sufficiently large so that there are sufficient number of columns having more than $t_0 := \lceil 2\mu_0^2 \log(2s_0\ell_0 n/\delta_0) \rceil$ overlaps with each seeds.
- This is again by Chernoff bound on the binomial distribution.

$$\gamma_0 \geq \sum_{j=t_0}^{\eta_0} \binom{\eta_0}{j} p_0^j (1-p_0)^{\eta_0 - j} \,. \quad \mathbb{P}(\widetilde{n} \leq \gamma_0 N/2) \leq \exp(-\gamma_0 N/8)$$

$$N \geq 2\ell_0 \gamma_0^{-1} n$$

$$N \geq \ell_0 n (2s_0\ell_0 n/\delta_0)^{2\mu_0^2 \log p_0^{-1}}$$

Step 2: Local matrix completion

- Given n by n matrix of rank r. This is simple matrix completion, except for three problems:
  1. Non-uniform sampling: a "thinning" operation
  2. Probability to complete the matrix for all seeds: Union bound
  3. There may be some neighborhood having columns from more than one subspaces.

"Thinning" operation as a fix for non-uniform sampling

- Local neighborhood is selected to be around the seed, so sampling is biased towards the support of the seed (denote by *t*)
- Key observations:
  - Those entries outside the *t* are not affected
  - Entries inside *t* with overlap *t'* has a probability to be greater than selection criteria *q* even if we are choosing randomly:
  - Prob(t'>= q) = $\rho = \sum_{j=q}^{t} \binom{t}{j} p_0^j (1-p_0)^{t-j}.$

“Thinning” operation as a fix for non-uniform sampling

- So thinning is conducted with two random variables *Y* and *Z*.
  - *Y = Bernoulli(ρ)*

$$\mathbb{P}(Z = j) = \frac{\binom{t}{j} p_0^j (1 - p_0)^{t-j}}{1 - \rho}$$

  - Define number of overlaps after thinning *t’’*

$$t'' = t'Y + Z(1 - Y)$$

$$\mathbb{P}(t'' = j) = \begin{cases} \mathbb{P}(Z = j)(1 - \rho) & j = 0, \ldots, q - 1 \\ \mathbb{P}(t' = j)\rho & j = q, \ldots, t \end{cases}$$

Guarantee successful local completion

- Lemma 7: Assume all s_0 matrices are "thinned", if sample rate satisfies that in the main theorem:

$$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \frac{r\,\log^2(n)}{n}$$

Then with probability $\geq 1 - 12 s_0 n^{2-2\beta^{1/2}} \log n$,

all s_0 *matrices can be perfectly completed.*

  - This probability is a relaxation from the probability of Simpler MC bound of Recht.
  - Assumption in A1: $k=o(n^d)$ for some d is used.

Matching sampling schemes (Lemma 7)

- This paper: iid Bernoulli Sampling
- Recht's Simpler MC paper: Uniform sampling with replacement.
- Solution:
  - Relax by a constant and show the condition holds with high probability
  - Turns out the condition is on the number of subspaces: as long as $k=o(e^n)$, the matching scheme succeeds with high probability.

Matching sampling schemes (Lemma 7)

- Each neighborhood has n^2 entries. Total number of samples follows Binomial. By Chernoff's Bound: $\mathbb{P}(\widehat{m} \leq n^2 p_0/2) \leq \exp(-n^2 p_0/8)$

$$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \frac{r\,\log^2(n)}{n} \qquad m' \geq 64 \max\left(\mu_1^2, \mu_0\right) \beta r n \log^2(2n)$$

- This \hat{m} is greater than requirement m', with high probability.
- By union bound, probability is multiplied by s_0

$$s_0 = O(k(\log k + \log n))$$

Columns from other subspace

- Lemma 5 guarantees with high probability, there are at least one seed for each subspace whose neighborhood is exclusively within the subspace. So correct subspaces are within all the completed subspace.
- Wrong subspace are span of multiple correct subspaces.

Step3: Subspace refinement

- Get all subspaces even if k is not known in prior.
- Sort the subspaces in increasing order of their dimension. Iterate over all subspaces and discard all that is contained in the union of subspace.

Step4: Full matrix completion

- Identifying subspace membership and recover full observation.
- Possible due to incoherence property.
- Why is it log(n), instead of log(N) of the typical matrix completion result?

Restricted subspace/shortened data vector

- Observation set: $\Omega \subset \{1, \ldots, n\}$
- U_Ω= is the restriction of rows of U to index Ω.
- Projection operator is naturally:

$$P_{S_\Omega} = U_\Omega \left(U_\Omega^T U_\Omega\right)^{-1} U_\Omega^T \ .$$

- The question is, how does the restricted residual reflect the total residual?

Restricted subspace/shortened data vector

- In k-GROUSE paper, and earlier In the "Matching subspace" paper
  - Assume incoherence μ(S), μ(y) and if number of observations is larger than Cμ(S)r log(r/δ) then with probability 1- δ

$$\frac{m(1-\alpha) - r\mu(S)\frac{(1+\beta)^2}{(1-\gamma)}}{n}\|v - P_S v\|_2^2 \leq$$

$$\|v_\Omega - P_{S_\Omega} v_\Omega\|_2^2 \leq (1+\alpha)\frac{m}{n}\|v - P_S v\|_2^2$$

  - α, β, γ are functions of μ, m, δ

$$\text{Let } v = x + y, \text{ where } x \in S \text{ and } y \in S^\perp.$$

Restricted subspace/shortened data vector

- Identify subspace membership of columns

Theorem 2. *Let* $\delta > 0$ *and* $m \geq \frac{8}{3} d_1 \mu(S^1) \log\left(\frac{2d_1}{\delta}\right)$. *Assume that*

$$\sin^2(\theta_0) < C(m) \sin^2(\theta_1) \ . \tag{7}$$

Then with probability at least $1 - 4\delta$,

$$\|v_\Omega - P_{S^0_\Omega} v_\Omega\|_2^2 < \|v_\Omega - P_{S^1_\Omega} v_\Omega\|_2^2 \ .$$

- Here the columns are perfect, condition (7) holds trivially. The rest also holds trivially since LFS is always 0, RHS is not 0 with high probability.

Subspace classification(Lemma 8)

- Lemma 8: Column x belonging to S1, partially observed on Ω can be classified using restricted projection $P_{\Omega,\mathcal{S}_j} = U_\Omega^j \left(\left(U_\Omega^j \right)^T U_\Omega^j \right)^{-1} \left(U_\Omega^j \right)^T,$

  If A3 holds and iid Bernoulli sampling with

$$p_0 \geq \frac{128\,\beta \max\{\mu_1^2, \mu_0\}}{\nu_0} \frac{r\,\log^2(n)}{n}$$

Then with probability at least $1 - (3(k-1)+2)\delta_0$,

$$\|x_\Omega - P_{\Omega,\mathcal{S}_1} x_\Omega\|_2^2 = 0$$

and for $j = 2, \ldots, k$

$$\|x_\Omega - P_{\Omega,\mathcal{S}_j} x_\Omega\|_2^2 > 0\,.$$

To wrap up the proof

- In the end, the full column is recovered by:

$$\widehat{x} = U \left(U_\Omega^T U_\Omega \right)^{-1} U_\Omega^T x_\Omega .$$

  with the same probability of Lemma 8.
- Take union bound of all probability of the previous development, the probability becomes: $1-(6+3(k-1)+12s_0)\delta_0 < 1-(6+15s_0)\,\delta_0$
- The proof is hence complete.

For full matrix completion

- A union bound over all N columns introduces the additional log(N) term on sample probability p_0.

- Alternatively, we may apply low rank matrix completion on all k subspaces, then apply union bound to get a better result: $O(knr\log^2(N/k))$

Reiterate the main points

- This paper proposed a method and theoretical guarantee for “High Rank” matrix completion problem.
- The proof largely relies on probabilistic argument and the assumptions are rather restrictive.
- All by itself, the sample rate is near optimal, but the matrix size must be very skewed to facilitate such subspace detection property.

Reiterate the main points

- General “high rank” matrix completion remains an open question.
- Specifically, it is an chicken egg problem of subspace clustering and matrix completion.
  - If full data is known, then subspace clustering is provably possible via SSC.
  - If subspace membership is known, then the data can be completed subspace by subspace via low-rank matrix completion.

Possible extensions?

- Relating to our own research, it is of great interests to propose alternative method subspace clustering algorithm with partial data.
- A possible extension is to add sample operator to either SSC or LowRank Representation (LR). The problem however, becomes non-convex and difficult to analyze.

Possible extensions?

- An alternative approach is the k-subspace clustering (analog to k-means) as described in the "k-subspace" paper of the same group.
- They proved that at each step, partial distance is almost the same as unknown full distance. So if any aspects of k-means are proved before (which there are some!) we can extend them to k-subspace with missing data.

Possible extensions?

- Other methods? Be creative!
- A convex formulation? Great!
- The nearest neighbor based method here is non-convex, yet provable!

Questions?