HIGH-DIMENSIONAL ESTIMATION WITH GEOMETRIC CONSTRAINS: REPORT

PARNIAN KASSRAIE

Introduction

Briefly explained, This paper introduces a simple and general estimator of a x, where x can be any vector or signal with any geometric constraints, coming from a model f, where f can be a nonlinear semi-parametric single-index model. The problem can include additive or non-linear noise.
Later an upper error bound is proven for their estimator, with a nice geometric interpretation for the error and minimum number of samples need for controlling that error up to a constant value.
Min-max optimality of the estimator for linear and nonlinear models is discussed under certain geometric condition and for specific fs, again, with a neat geometric analogy.

This report with include a concise summery of the paper as well as a detailed section on my contribution that was presented in class.

> The Reader can easily skip Paper Summery (section 1) and start reading from the Matrix Completion (section 2) which is my additional contribution to the paper.

1. Paper Summery

1.1. **The Model & It's Upper Error Bound.** The following model is used throughout the paper:

$$y_i = f(\langle a_i, x\rangle + \epsilon), \quad 1 \le i \le m$$

Where $f : \mathbb{R} \to \mathbb{R}$ is an unknown function, $x \in \mathbb{R}^n$ is the fixed unknown signal and $a_i \sim \mathcal{N}(0, I_n)$ are i.i.d random variables, the observation or sensing vectors. Therefore,

> each observation y_i may depend on a_i only through $\langle a_i, x\rangle$

In our problems, often we have a prior knowledge of x. This information is modeled by saying $x \in K$. Further on, we will assume that K is a Star-shaped ($\forall\ 0 \le \lambda \le 1 : \lambda K \subseteq K$) subset of $\mathbb{R}^n$.

Now let's see how we can estimate x.

1.1.1. ***Linear Estimator.*** The simplest approach is to ignore K, which is common in ML estimations. The linear estimator is then defined as:

$$\hat{x}_{lin} := \frac{1}{m}\sum_{i=1}^{m} y_i a_i$$

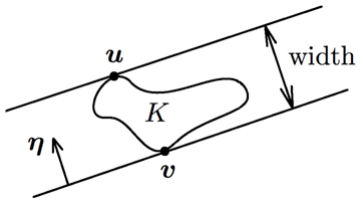

FIGURE 1. Mean Width Illustration

It can be seen that it's an unbiased estimator of normalized x. For the error bound we have the following theorem:

Theorem 1.1.

$$\mathbb{E}\hat{x}_{lin} = \mu\bar{x}, \quad \sqrt{\mathbb{E}\left\|\hat{x}_{lin} - \mu\bar{x}\right\|_2^2} = \frac{1}{\sqrt{m}}[\sigma + \eta\sqrt{n}]$$

Where

$$\bar{x} = \frac{x}{\|x\|_2}, \quad \mu = \mathbb{E}y_1 < a_1, \bar{x} >, \quad \sigma^2 = Var(y_1 < a_1, \bar{x}), \quad \eta^2 = \mathbb{E}y_1^2$$

Proof can be found in the paper, section 8.2.

Now let's explore the estimator introduced in the paper, but before, we have to get familiar with concept of mean width.

1.1.2. ***Mean Width***. The width of K in the direction of a given unit vector $\eta \in S^{n-1}$ is defined as the width of the smallest slab between two parallel hyperplanes with normals η that contains K; Figure 1.We can express the width in the direction of as

$$\sup_{u,v\in K} \langle \eta, u - v\rangle = \sup_{z\in K-k} \langle \eta, z\rangle$$

where $KK = \{uv : u, v \in K\}$ is the Minkowski sum of K and K. Averaging over η uniformly distributed on the sphere S^{n1}, we can define the spherical mean width of K:

$$\tilde{w}(K) := \mathbb{E} \sup_{z\in KK} \langle \eta, z\rangle$$

Definition 1.2 (Gaussian Mean Width)**.** The Gaussian mean width of a bounded subset K of $\mathbb{R}^n$ is defined as

$$w(K) := \mathbb{E} \sup_{u\in KK} \langle g, u\rangle$$

where $g \sim \mathcal{N}(0, I_n)$ is a standard Gaussian random vector in $\mathbb{R}^n$. We will refer to Gaussian mean width as simply the mean width.

Local Mean Width is often used in this paper's theorems, so we'll quickly define it as well:

$$w_t(K) = \mathbb{E}_g \sup_{x,y\in K} \langle g, x - y\rangle, \quad \|x - y\|_2 \le t$$

Some examples of famous mean widths:

- B_2^n:

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

- A subset of B_2^n with linear algebraic dimension d:

$$w(K) = 2\sqrt{d}$$

- A finite subset of B_2^n:

$$w(K) = C\sqrt{|K|}$$

- $\Sigma_k \cap S^{n-1}$:

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

1.1.3. ***Projected Estimator & Error Bound***. The straight forward step took in the paper towards improving the linear estimator, is simply projecting it onto the feasible set K.

$$\hat{x} = P_K(\hat{x}_{lin}) = \underset{z \in K}{\operatorname{argmin}} \|\hat{x}_{lin} - z\|_2$$

The main result of the paper is the upper error bound for projected estimator:

Theorem 1.3. *(Main Result) For every $t > 0$:*

$$\mathbb{E}\,\|\hat{x} - \mu\bar{x}\| \leq t + \frac{2}{\sqrt{m}}\left[\sigma + \eta\frac{w_t(K)}{t}\right]$$

Optimizing equation above on t:

$$\mathbb{E}\,\|\hat{x} - \mu\bar{x}\| \leq \frac{2\sigma}{\sqrt{m}} + 2\sqrt{2}\left[\eta\frac{w(K)}{\sqrt{m}}\right]^{1/2}$$

Where μ, η, σ are defined similar to the linear case.

The theorem shows that in order to have a constant error rate, the sample size (m) should be as large as mean width of K squared. Which is a nice reduction for the inherently sparse feasible sets.
In the following we will see a couple of examples for better understanding of this abstract model.
Examples of common in practice feasible sets:

- Sparse Vectors

$$K = \Sigma_s$$

$$w(K) \leq C\sqrt{s \log \frac{2n}{s}}$$

$$m = O(s \log \frac{2n}{s})$$

- Approximately Sparse Vectors

$$K = \{v \in \mathbb{R}^n : \|v\|_1 \leq s\}$$

$$w(K) \leq 4\sqrt{2s \log n}$$

$$m = O(s \log n)$$

Examples of common in practice $f-$Models:

- Noisy Linear Model

$$y_i = \langle a_i, x\rangle + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \nu^2)$$

The error bound would become:

$$\mathbb{E}\,\|\hat{x} - \mu\bar{x}\| \leq t + \frac{\|x\|_2 + v}{\sqrt{m}}\left[1 + \frac{w_t(K)}{t}\right]$$

The power of this could be seen when the noise energy is comparable to the signal ($\|x\|_2 + v$ term), the error rate will not increase dramatically and the estimator, to some extent, would still be valid.

- Non-linear Model

$$y_i = f(\langle a_i, x\rangle)$$

If we calculate η, σ for this setting, we would have:

$$\sigma^2 = Var\big[f(g\,\|x\|_2)g\big], \quad \eta^2 = \mathbb{E}f(g\,\|x\|_2)^2$$

Where $g \sim \mathcal{N}(0,1)$. Recalling the main theorem:

$$\mathbb{E}\,\|\hat{x} - \mu\bar{x}\| \leq t + \frac{2}{\sqrt{m}}\left[\sigma + \eta\frac{w_t(K)}{t}\right]$$

For heavy-tail nonlinear f, we will thus have a looser error bound.

1.2. **Model Optimality.** Projected estimator error bound is compared with a general lower bound for **any estimator** that is only a function of y_i, a_i. By comparing these bounds, to sufficient number of samples in order to have a minimax optimal error is calculated. Meaning, having this many samples, there exists no other estimator that could perform orders of magnitude better than the projected estimator, because the projected estimator's upper error bound is controlled by the general lower bound up to a constant factor.

Packing Number. In order to understand the lower error bound we need to introduce the following notation.

Definition 1.4. (Packing Number) Packing number of K with balls of radius t:

$$\mathbb{P}_K = \sup_{\chi} |\chi|$$

Where,

$$\chi \subset K : \ \|v - w\|_2 \geq t, \quad \forall v, w \in \chi.$$

Figure 2 can give a geometric understanding of the concept.

Similar to mean width, Local Packing Number is defined as:

$$\mathbb{P}_t = \mathbb{P}_{K \cap tB_2^n}$$

with balls of radius $t/10$. A useful parameter is local packing number to local mean width ratio which will be used later:

$$\alpha = \alpha(K) = \sup_t \frac{w_t(K)}{t\sqrt{\log P_t}}$$

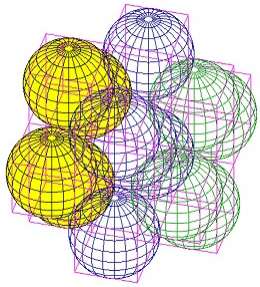

FIGURE 2. Pakcing Number Illustration

1.2.1. ***Linear Model Optimality***.

Theorem 1.5. *(Linear Model Lower Error Bound) For Any $\hat{x}$ which depends only on y_i, a_i*

$$\delta_* := \inf_{t>0}\{t + \frac{\nu}{\sqrt{m}}[1 + \sqrt{\log P_t}]\}$$

$$\sup_{x\in K} \mathbb{E}\,\|\hat{x} - x\|_2 \geq c \min\big(\delta_*, diam(K)\big)$$

We The main result theorem can be re-written in a way that is similar to the lower bound theorem above:

$$\mathbb{E}\,\|\hat{x}_{proj} - x\|_2 \leq C \inf_t\{t + \frac{\nu + \|x\|_2}{\sqrt{m}}\big[1 + \frac{w_t(K)}{t}\big]\} = C\delta^*$$

$$\sup_{x\in K} \|\hat{x}_{proj} - x\|_2 \leq c \min\big(\delta^*, diam(K)\big)$$

It can be easily seen that projected estimator for linear models is optimal if:

$$\delta_* \geq \delta^*/\alpha$$

Where α is the radio defined in the previous section. Therefore, for all feasible sets that $\alpha(K)$ is controlled by a constant value the model is optimal, thus a great choice among all estimators. I must mention that such feasible sets aren't hard to find, sparse vectors, low-rank matrices and approximately sparse vectors are some familiar examples.
In addition Near-Optimality holds for non-linear models in special cases of f, K which is skipped here since it wasn't included in the presentation.

2. An Application Extending the Paper: Matrix Completion

2.1. **The Matrix Completion Problem.** As a practical addition to the paper, I decided to apply the Projected Estimator on Low-Rank Matrix Completion Problem.

Low-rank Matrix $X \in \mathbb{R}^{d_1\times d_2}$, $rank(X) \leq r$ is partially observed. The index of observed cells are members of $\Omega \subset \{1, \cdots, d_1\} \times \{1, \cdots, d_2\}$.
We define Δ as the observation mask matrix,

$$(\Delta)_{ij} = \mathbb{1}_{(i,j)\in\Omega}$$

Thus, The observed matrix can be formulated as:

$$X_{obs} = X \circ \Delta$$

Where $\circ$ is the Hadamard (Entry-wise) product of the two matrices.

Now let's see the estimator this paper suggests for estimating X.

2.2. **Plan's Suggested Solution.** The linear unbiased estimator for X is $\hat{X}_{lin} = X_{obs}$. However, we know that X is low-rank. Thus, $X \in K$ where

$$K = \{M \in \mathbb{R}^{d_1 \times d_2} : rank(M) \leq r\}$$

Therefore, Plan's Estimator would be:

$$\hat{X} = P_k(\hat{X}_{lin})$$

Basically, we should simply project the observed matrix onto the set of low-rank matrices. The estimation error, $\left\|X - \hat{X}\right\|_2$ is bounded the squared mean width of this set, which is $W^2(K) \leq 2r(d_1 + d_2)$.

Issues. As you might have noticed, projecting X_{obs} onto K would change the value of observed cells. In many of the applications, e.g. Image Reconstruction, preserving the observed information is a necessity. Thus, Plan's Estimator wouldn't be a wise option here.

Consistency is an important feature of an estimator, which has received no attention throughout the paper. As explained above, the suggested paper is an unbiased estimator that will not converge to X in the asymptotic setting of,

$$d_1, d_2, r \to \infty,\ \frac{r}{d_1}, \frac{r}{d_2} \in [0, 1]$$

as long as $r < \min(d_1, d_2)$. This really is an unpleasant property!

2.3. **The Alternative Solution.** A simple solution to the issue above, would be projecting $\hat{X}$ onto W,

$$W = \{M \in \mathbb{R}^{d_1 \times d_2},\ M_\Omega = X_\Omega\}$$

Which is an Affine space containing the matrices with correct cells at the observed indices. Projection onto W however, might push the estimator out of K. We should look for matrices inside $K \cap W$.

Main Idea. We keep projecting $\hat{X}_{lin}$ onto K, W repeatedly, for as many times as desired, finishing with a projection to W. The method is illustrated in Figure 3. In the following I will show how this method improves the original estimator in theory and in practice.

2.3.1. ***Solving the Consistency Issue.*** Assuming that $K \cap W \neq \emptyset$, since the real $X \in K \cap W$, the alternative method will tend to find it. In case of $K \cap W = \emptyset$ this method will try to find $\hat{X}_{new}$, where

$$\hat{X}_{new} = \underset{M \in W}{\operatorname{argmin}} \{ \min_{N \in K} \|M - N\| \}$$

The following proposition will prove consistency of this estimator in the cases where both W, K are convex set. Notice that in our example K is not convex.

Theorem 2.1. *For compact convex spaces A, B, where $A \cap B \neq \emptyset$, By starting from a given point x, and periodically projecting it onto A and B, the sequence of resulting points $\{x_i\}$ will converge to a point inside $A \cap B$, where $\{x_i\}$ can be formulated as:*

$$x_0 : x,\ x_1 : P_A(x_0),\ x_2 : P_B(x_1),\ x_3 : P_A(x_2), \cdots$$

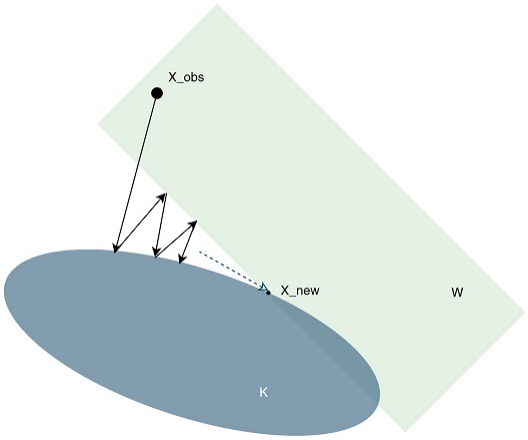

FIGURE 3. Alternative Method

Lemma 2.2. *Given $x, y \in \mathbb{R}^n$ and the compact convex set C we have:*

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|$$

Where $\|.\|$ is the l_2 norm.

The proof is skipped since it could be easily derived or found online.

Definition 2.3. Distance of a point from a set is defined as:

$$d(x, C) = \min_{y \in C} \|x - y\|$$

Lemma 2.4. *For $A, B, \{x_i\}$ defined in Theorem 2.2, the sequence $\alpha_n = d(x_n, A \cap B)$ is decreasing.*

Proof. Without a loss of generality, assume that $x_{n+1} = P_A(x_n)$. We define $y = P_{A \cap B}(x_n)$, But $y \in A$ hence $P_A(y) = y$. From Lemma 2.3:

$$d(x_{n+1}, A \cap B) \leq \|x_{n+1} - y\| \leq \|x_n - y\| = d(x_n, A \cap B)$$
$$d(x_{n+1}, A \cap B) \leq d(x_n, A \cap B)$$

□

The only step left to prove Theorem 2.2 is to show convergence of $d(x_n, A \cap B)$ to zero.

Proof Theorem 2.2.

$$\exists\, \epsilon \text{ s.t. } \forall n \in \mathbb{N}\ d(x_n, A \cap B) > \epsilon$$

Given convex compact set C and point x outside C:

$$P(x) = \operatorname*{argmin}_{y \in C} \|y - x\|^2$$
$$\Rightarrow \angle y P(x) x \geq \pi/2$$
$$\Rightarrow \|x - y\|^2 \geq \|x - P(x)\|^2 + \|y - P(x)\|^2$$

The equations above are illustrated in Figure 4.
Therefore if the contradictory assumption holds, we propose that for a sample point x^* in $A \cap B$ the sequence $\|x_n - x^*\|$ is decreased by δ^2 with every projection. Where

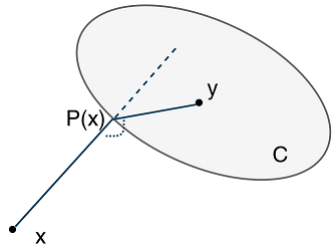

FIGURE 4. Theorem 2.2 Proof

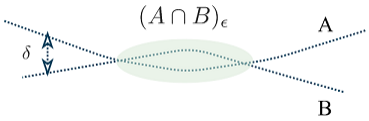

FIGURE 5. Theorem 2.2 Proof

δ is the shortest distance between $A/(A \cap B)_\epsilon$ and B or reverse.
On each step x_i is either inside $A/(A \cap B)_\epsilon$ or $B/(A \cap B)_\epsilon$ and is being projected on B or A respectively. Assume the first case, since $A/(A \cap B)_\epsilon \cap B = \emptyset$ then $d(x_i, B) \geq \delta$ thus, $d(x_i, A \cap B)$ is reduced by δ^2. (Check out Figure 5.) The opposite hypothesis is not true and the proposition is proven:

$$d(x_n, A \cap B) \xrightarrow{n \to \infty} 0$$

The last step is to prove the sequence $\{x_i\}$ converges to a point inside $A \cap B$, which is done by showing it's Cauchy sequence.
For $\epsilon > 0$, $\exists x_n$, $x^* \in A \cap B$ that $\|x_n - x^*\| \leq \epsilon$, thus $\forall m \geq n : \|x_m - x^*\| < \epsilon$. Therefore, all the following members of the sequence are inside a ball centered at x_n with radius ϵ and the sequence is Cauchy.

Convergence of $\{x_i\}$ to $A \cap B$ is thus proven. □

2.3.2. ***Implementation.*** It's a well-known fact that images are low-rank matrices, thus the estimator is implemented in MATLAB as a method for partially observed image reconstruction.
Projection onto the space of low-rank matrices is done by calculating SVD and keeping the $r-$largest eigenvalues i.e. $r-$largest cells on Σ's diagonal where $X = U\Sigma V^T$.
A few points about the simulations:

- The original image's dimensions are 854×1280.
- For $r \in \{3, 10, 50, 100\}$, each time $m = O(r(d_1 + d2))$ pixels are observed.
- The indices (Ω) are independent uniform random variables.
- The final reported SNRs are calculated by averaging on the SNRs of 10 different instances of Ω.
- The algorithm has 20 projection steps between K, W.

FIGURE 6. Original image selected for partial observations

Results. Original image is shown in Figure 6. Figure 7 shows the result of implementing Plan's estimator and Figure 8 shows the result of implementing the alternative estimator. The codes can be found in *ImageRecovery.m*.

I have chosen SNR as an indicator for quality of estimation. Table 1. shows how the alternative method improves Plan's estimator for larger assumed image matrix ranks.

TABLE 1. SNR Comparison for different assumed ranks

Rank	Plan's	Ours
3	−40.5565	−41.6303
10	1.0223	1.1253
50	0.9369	0.9292
100	2.9225	2.9389

A pivotal idea in the paper, is bounding the estimation error with mean width of K and thus, coming up with minimum number of observations needed for controlling the error up to a constant value. In Table 2 I have assumed $r = 50$ and then compared the average SNR for Plan's method with different orders of observation. As it can be seen, the suggested order(around 100 thousand samples) significantly

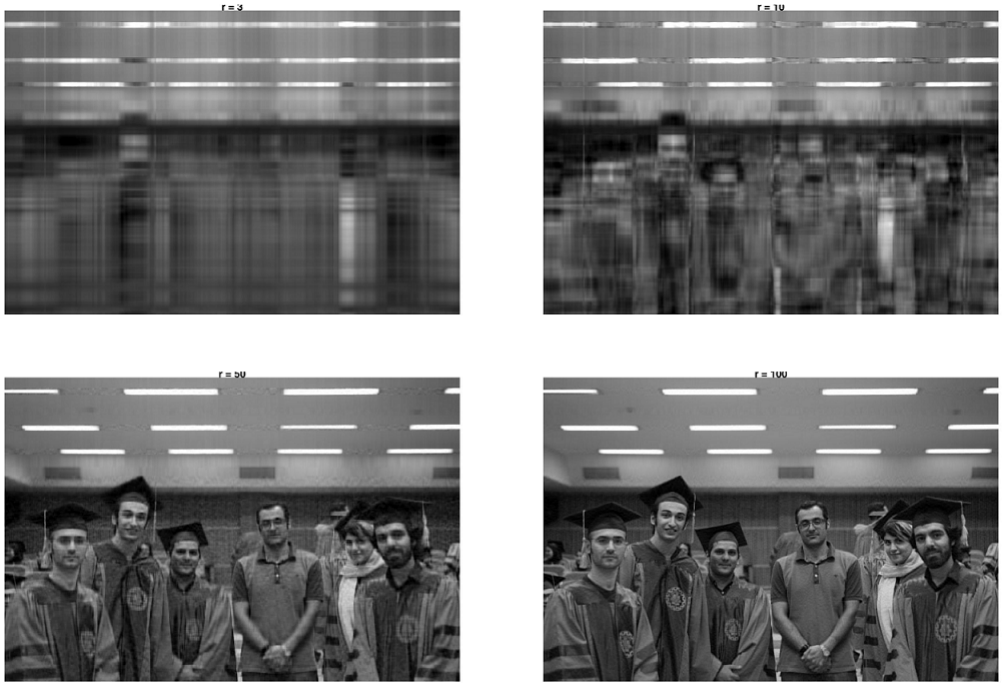

FIGURE 7. Projected Estimator for Low-rank Matrix Recovery

improves $O(1)$ and is almost as good as observing all the cell values (about 1 million samples).

TABLE 2. SNR Comparison for different number of observations, using Plan's estimator

order of m	Avg. SNR
Const. Rate	-50.2226
$r(d_1+d_2)$	0.9369
$d_1 d_2$	0.9409

Further Improvements. The astute reader might have noticed that the proven theory was only on convex compact sets, while we are dealing with non-convex set of low-rank matrices in our problem formulation and implementations. A nice way to remove this problem would be using a convex relaxation of K:

$$\tilde{K} = \{M \in \mathbb{R}^{d_1 \times d_2} : \|M\|_* \leq r\}$$

Where $\|.\|_*$ is the matrix nuclear norm. (It is only defined for positive definite matrices, but well.) This method was simultaneously introduced by M. Fazel in 2002 and later extended by Candes & Tao in 2010.

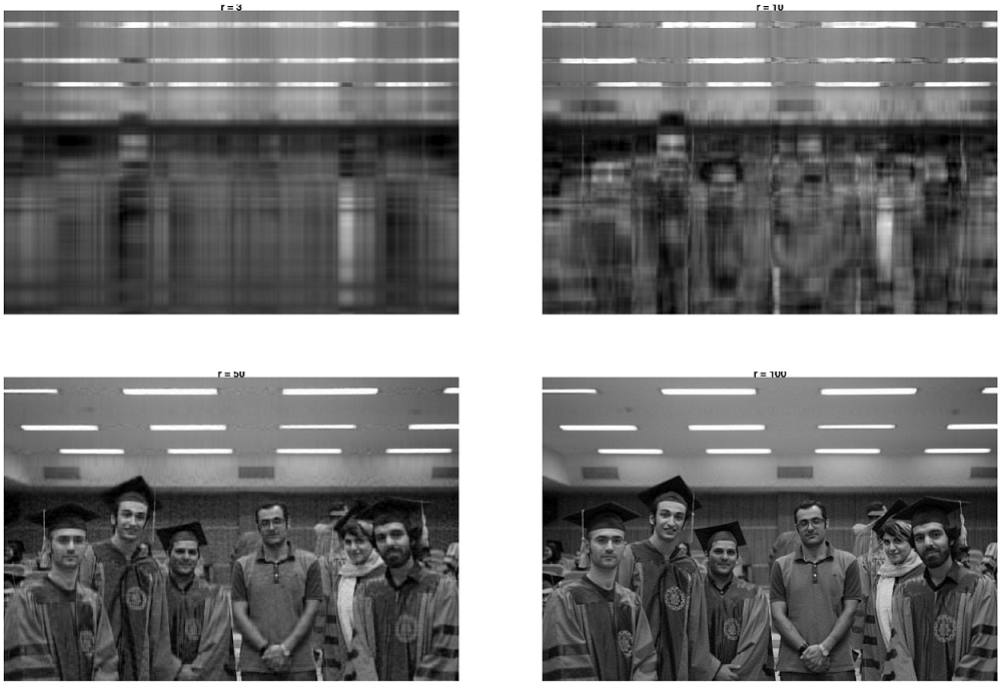

FIGURE 8. Projected Estimator for Low-rank Matrix Recovery

REFERENCES

1. A. A. Giannopoulos, V. D. Milman *Asymptotic Convex Geometry Short Overview." Different faces of geometry.* Different Faces of Geometry, Springer, Boston, MA, 2004. 87-162.
2. R. Vershynin, *Estimation in High Dimesntions: A Geometric Perspective.* Sampling theory, a renaissance. Birkhuser, Cham, 2015. 3-66.
3. Y. Plan, R. Vershynin *"Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach.*, IEEE Transactions on Information Theory 59.1 (2013): 482-494.
4. K. Alishahi, H. Foroughmand *Convex Optimization*, Sharif University of Technology, Spring 2018.
5. K. Alishahi, *High-Dimensional Data Analysis*, Sharif University of Technology, Fall 2017.
6. B. Recht, M. Fazel and A. Parrilo *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*2010, SIAM review, 52(3), pp.471-501.
7. E.J. Cands, T. Tao *The power of convex relaxation: Near-optimal matrix completion.*, 2010, IEEE Transactions on Information Theory, 56(5), pp.2053-2080.

COMPRESSED SENSING, SHARIF UNIVERSITY OF TECHNOLOGY, SPRING 2018
E-mail address: kasraei_parnian@ee.sharif.edu