Guaranteed Rank Minimization via Singular Value Projections

Inderjit S. Dhillon
University of Texas at Austin

Workshop on Algorithms for Massive Data Processing
IIT Kanpur
Dec 18, 2009

Joint work with Raghu Meka, Prateek Jain

Overview

- Affine Constrained Rank Minimization Problem (ARMP)
- Singular Value Projection algorithm (SVP)
- Analysis
- Matrix Completion
- Results
- Conclusions

Rank Minimization Problem(RMP)

$$
\begin{aligned}
(\textbf{RMP}) : \min \quad & \text{rank}(X) \\
\text{s.t} \quad & X \in \mathcal{C}.
\end{aligned}
$$

- $\mathcal{C}$ is a convex set, e.g., a polyhedral set
- Applications:
  - Machine Learning
  - Computer Vision
  - Control Theory

Affine Constrained Rank Minimization Problem (ARMP)

$$(\textbf{ARMP}): \quad \min_{X} \;\; rank(X)$$
$$\text{s.t} \;\; \mathcal{A}(X) = b.$$

$$X \in \mathbb{R}^{m \times n}, \;\mathcal{A}: \mathbb{R}^{m \times n} \to \mathbb{R}^{d}, \;\; b \in \mathbb{R}^{d}.$$

- $d \ll mn$
- Applications:
  - Matrix completion: Netflix Challenge
  - Linear time-invariant systems
  - Embedding using missing Euclidean distances
- NP-hard even to approximate within log factor (Meka et al.'08)

An Example: Minimum Rank Matrix Completion

- Netflix Challenge:
  - Given a few user-movie ratings
  - Goal: complete ratings matrix
- Small number of latent factors $\equiv$ low-rank
- Special case of ARMP:

$$(\mathbf{MCP}): \quad \min_{X} \quad rank(X)$$
$$\text{s.t} \quad \mathrm{tr}(X\mathbf{e}_j\mathbf{e}_i^T) = b_{ij}, \forall (i,j) \in \Omega.$$

- Typically, number of samples very small: Netflix has 1% samples

Existing Work

- Various heuristics like alternative minimization, log-det relaxation
- Typically no theoretical guarantees
- Recent work: theoretical guarantees from generalizations of compressed sensing

ARMP: Generalization of Compressed Sensing (CS)

$$(\mathbf{CS}): \quad \min_{\mathbf{x}} \quad \|\mathbf{x}\|_0$$
$$\text{s.t} \quad A\mathbf{x} = b.$$

$$\mathbf{x} \in \mathbb{R}^n, \; A : \mathbb{R}^n \to \mathbb{R}^d, \; b \in \mathbb{R}^d.$$

- $d \ll n$ (typically, $d = s \log n$)
- Specific instance of ARMP with $X = Diag(x)$.

Technique	CS	ARMP
Convex relaxation	ℓ_1 (Lasso)	Trace-norm (SVT)
Greedy approach	MP, OMP, CoSamp	ADMiRA
Hard Thresholding	IHT, GradeS	SVP, IHT

Table: CS vs ARMP

Restricted Isometry Property (RIP)

- Most CS methods assume RIP

$$(1-\delta_s)\|\mathbf{x}\|^2 \le \|A\mathbf{x}\|^2 \le (1+\delta_s)\|\mathbf{x}\|^2,\ \forall \mathbf{x} \text{ s.t. } \|\mathbf{x}\|_0 \le s$$

- Generalization to matrices:

$$(1-\delta_k)\|X\|_F^2 \le \|\mathcal{A}(X)\|_2^2 \le (1+\delta_k)\|X\|_F^2,\ \forall X \text{ s.t. } \mathrm{rank}(X) \le k$$

- Families satisfying RIP:

$$\mathcal{A}(X) = A\ vec(X),$$

  - $A_{ij} \sim \mathcal{N}(0, 1/d)$
  - $A_{ij} = \begin{cases} 1/\sqrt{d} & \text{with probability } 1/2 \\ -1/\sqrt{d} & \text{with probability } 1/2 \end{cases}$

Singular Value Projection (SVP)

$$(\textbf{RARMP}) : \min_{X} \psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2,$$
$$\text{s.t } X \in \mathcal{C}(k) = \{X : rank(X) \leq k\}.$$

- Adapt classical projected gradient
- Efficient projection onto non-convex rank constraint

Singular Value Projection (SVP)

Algorithm 1 SVP algorithm

Initialize $X^0 = 0,\ t = 0$
Set step size η_t
repeat
$$X^{t+1} = P_k(X^t - \eta_t \underbrace{\mathcal{A}^T(\mathcal{A}(X^t) - b)}_{\nabla\psi(X)})$$
$t = t + 1$
until Convergence

- $P_k(X) = U_k \Sigma_k V_k^T$—top k singular vectors, best rank k approximation
- X^t: low-rank, stored using $(m+n)k$ values

SVP: Main result

Theorem

Isometry constant: $\delta_{2k} < 1/3$
Exact case: $b = \mathcal{A}(X^*)$
Set $\eta_t = 1/(1 + \delta_{2k})$
SVP outputs matrix X *of rank* k *s.t.*

$$\|\mathcal{A}(X) - b\|_2^2 \leq \epsilon$$

Maximum number of iterations:

$$\left\lceil C \log \frac{\|b\|^2}{2\epsilon} \right\rceil$$

- Geometric convergence
- For $\delta_{2k} = 1/5$, $\eta_t = 5/6$, number of iterations: $\left\lceil \log_2 \frac{\|b\|^2}{2\epsilon} \right\rceil$

SVP: Guarantees—Noisy Case

Theorem

Isometry constant $\delta_{2k} \leq 1/3$
Noisy case: $b = \mathcal{A}(X^*) + e$ *(e is error vector)*
Set $\eta_t = 1/(1 + \delta_{2k})$
SVP outputs X *of rank* k *s.t.,*

$$\|\mathcal{A}(X) - b\|_2^2 \leq (C^2 + \epsilon)\|e\|^2, \ \epsilon \geq 0$$

Number of iterations is bounded by:

$$\left\lceil D \log \frac{\|b\|^2}{(C^2 + \epsilon)\|e\|^2} \right\rceil$$

- Geometric convergence to C-approx solution

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) = \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\|\mathcal{A}(\overbrace{X^{t+1} - X^t}^{\text{Rank}2k})\|_2^2$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) = \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\|\mathcal{A}(\overbrace{X^{t+1} - X^t}^{\text{Rank}2k})\|_2^2$$
$$\leq \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\underbrace{(1 + \delta_{2k})\|X^{t+1} - X^t\|_F^2}_{\text{Using RIP}},$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) = \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\|\mathcal{A}(\overbrace{X^{t+1} - X^t}^{\text{Rank}2k})\|_2^2$$

$$\leq \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\underbrace{(1+\delta_{2k})\|X^{t+1} - X^t\|_F^2}_{\text{Using RIP}},$$

$$= \frac{1}{2}(1+\delta_{2k})\|X^{t+1} - Y^{t+1}\|_F^2 - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) = \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\|\mathcal{A}(\overbrace{X^{t+1} - X^t}^{\text{Rank}2k})\|_2^2$$

$$\leq \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\underbrace{(1+\delta_{2k})\|X^{t+1} - X^t\|_F^2}_{\text{Using RIP}},$$

$$= \frac{1}{2}(1+\delta_{2k})\|X^{t+1} - Y^{t+1}\|_F^2 - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

$$where Y^{t+1} = X^t - \frac{1}{1+\delta_{2k}}\nabla\psi(X^t), \; X^{t+1} = P_k(Y^{t+1})$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) = \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\|\mathcal{A}(\overbrace{X^{t+1} - X^t}^{\text{Rank}2k})\|_2^2$$

$$\leq \langle \nabla\psi(X^t), X^{t+1} - X^t \rangle + \frac{1}{2}\underbrace{(1+\delta_{2k})\|X^{t+1} - X^t\|_F^2}_{\text{Using RIP}},$$

$$= \frac{1}{2}(1+\delta_{2k})\|X^{t+1} - Y^{t+1}\|_F^2 - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

$$\leq \frac{1}{2}(1+\delta_{2k})\underbrace{\|X^* - Y^{t+1}\|_F^2}_{Eckart-Young Theorem} - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) \leq \frac{1}{2}(1+\delta_{2k}) \underbrace{\|X^* - Y^{t+1}\|_F^2}_{Eckart-Young Theorem} - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

$$= \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2}(1+\delta_{2k})\|X^* - X^t\|_F^2$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) \leq \frac{1}{2}(1+\delta_{2k}) \underbrace{\|X^* - Y^{t+1}\|_F^2}_{Eckart-Young Theorem} - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

$$= \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2}(1+\delta_{2k})\|X^* - X^t\|_F^2$$

$$\leq \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2} \underbrace{\frac{1+\delta_{2k}}{1-\delta_{2k}}\|\mathcal{A}(X^* - X^t)\|_2^2}_{\text{Using RIP}}$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) \leq \frac{1}{2}(1+\delta_{2k}) \underbrace{\|X^* - Y^{t+1}\|_F^2}_{Eckart-Young Theorem} - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

$$= \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2}(1+\delta_{2k})\|X^* - X^t\|_F^2$$

$$\leq \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2} \underbrace{\frac{1+\delta_{2k}}{1-\delta_{2k}}\|\mathcal{A}(X^* - X^t)\|_2^2}_{\text{Using RIP}}$$

$$= \psi(X^*) - \psi(X^t) + \frac{\delta_{2k}}{(1-\delta_{2k})}\|\mathcal{A}(X^* - X^t)\|_2^2,$$

SVP: Proof

- Simple analysis–apply RIP twice and Eckart-Young theorem once
- $\psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$: a quadratic function,

$$\psi(X^{t+1}) - \psi(X^t) \leq \frac{1}{2}(1+\delta_{2k}) \underbrace{\|X^* - Y^{t+1}\|_F^2}_{Eckart-Young Theorem} - \frac{1}{2(1+\delta_{2k})}\|\mathcal{A}^T(\mathcal{A}(X^t) - b)\|_F^2$$

$$= \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2}(1+\delta_{2k})\|X^* - X^t\|_F^2$$

$$\leq \langle \nabla\psi(X^t), X^* - X^t \rangle + \frac{1}{2} \underbrace{\frac{1+\delta_{2k}}{1-\delta_{2k}}\|\mathcal{A}(X^* - X^t)\|_2^2}_{\text{Using RIP}}$$

$$= \psi(X^*) - \psi(X^t) + \frac{\delta_{2k}}{(1-\delta_{2k})}\|\mathcal{A}(X^* - X^t)\|_2^2,$$

For exact case, $\psi(X^*) = 0$, $\mathcal{A}(X^*) = b$. Hence,

$$\psi(X^{t+1}) \leq \underbrace{\frac{2\delta_{2k}}{(1-\delta_{2k})}}_{<1 \text{ for } \delta_{2k}<1/3} \psi(X^t).$$

Comparison to Existing Methods

Method	Generalization of	RIP constant	Rate of Convergence	Noisy Measurements
Trace-norm (RFP07)	l_1 relaxation	$\delta_{5k} < 1/10$	Not known	No
Trace-norm (LB09b)	l_1 relaxation	$\delta_{3k} < 1/4\sqrt{3}$	Not known	Yes
ADMiRA (LB09a)	Matching Pursuit	$\delta_{4k} < 1/\sqrt{32}$	Geometric	Yes
SVP	IHT	$\delta_{2k} \leq 1/3$	Geometric	Yes

Table: Comparison of the existing approaches with SVP

Matrix Completion

- Complete a low-rank matrix from few sampled entries
- Minimum rank matrix completion problem:

$$(\mathbf{MCP}) : \min_{X} \operatorname{rank}(X), \quad \text{s.t } \mathcal{P}_{\Omega}(X) = \mathcal{P}_{\Omega}(X^*).$$

- $\mathcal{P}_{\Omega} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$—projection onto index set Ω, i.e.,

$$(\mathcal{P}_{\Omega}(X))_{ij} = \begin{cases} X_{ij} & \text{for } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

- Special case of ARMP: SVP can be applied directly
- Problem: MCP does not satisfy RIP in general

Existing Work: Matrix Completion

- Most ARMP methods applicable to MCP
- Exact recovery:
  - Trace-norm relaxation: Recht and Candes'08, Candes and Tao'09
  - SVD+Alternative Minimization: Keshavan et al.'09
- Assumptions: uniform sampling, incoherence

Definition (Incoherence)

$X \in \mathbb{R}^{m \times n}$ with SVD $X = U\Sigma V^T$ is μ-incoherent if

$$\max_{i,j} |U_{ij}| \leq \frac{\sqrt{\mu}}{\sqrt{m}}, \quad \max_{i,j} |V_{ij}| \leq \frac{\sqrt{\mu}}{\sqrt{n}}.$$

SVP: Matrix Completion

$$(\mathbf{RMCP}) : \min_X \ \psi(X) = \frac{1}{2}\|P_\Omega(X - X^*)\|_F^2,$$
$$\text{s.t } X \in \mathcal{C}(k) = \{X : rank(X) \le k\}.$$

Algorithm 2 SVP for Matrix Completion

Initialize $X^0 = 0,\ t = 0$
Set step size $\eta_t = 1/(1+\delta)p$, p=sampling density, δ is a parameter
repeat
  $X^{t+1} = P_k(X^t - \eta_t P_\Omega(X^t - X^*))$
  $t = t + 1$
until Convergence

- $P_k(X) = U_k \Sigma_k V_k^T$: top k singular vectors of X
- Computation of k singular vectors of: $\underbrace{X^t}_{low\ \ rank} - \eta_t \underbrace{P_\Omega(X^t - X^*)}_{sparse}$
- Matrix-vector multiplication: $O((m+n)k + |\Omega|)$

Weak RIP

- We show the following weak RIP:

Theorem (Weak RIP)

Let $0 < \delta < 1$.
Sampling density $p \geq C\mu^2 k^2 \log n/\delta^2 m$.
For all rank k, μ-incoherent matrices X,

$$(1-\delta)p\,\|X\|_F^2 \;\leq\; \|\mathcal{P}_\Omega(X)\|_F^2 \;\leq\; (1+\delta)p\,\|X\|_F^2,$$

with high probability.

- Similar to RIP but only for incoherent matrices
- If every iterate is incoherent $\implies$ SVP optimal

Matrix Completion: Convergence Proof?

- SVP converges if:

Conjecture (Incoherence)

Let X and X^* be rank k, μ-incoherent matrices.
Set $\eta_t < 1$.

$$Y = P_k(X - \eta_t P_\Omega(X - X^*))$$

is $(1+\epsilon)\mu$-incoherent for small ϵ.

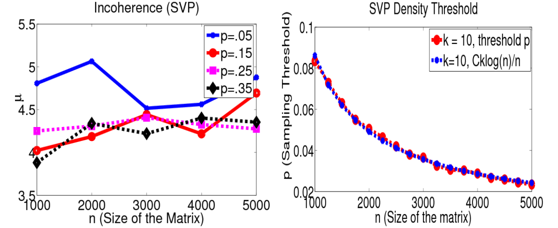


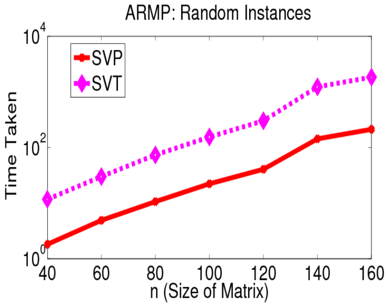
Figure: Empirical estimates of incoherence and sampling density threshold (matches $Ck \log n/n$, $C = 1.28$)

Results: ARMP

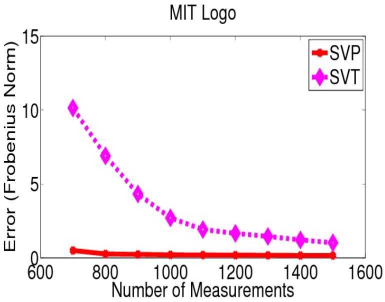
- Synthetic Datasets:
  - Generate a random X^* of rank k
  - Generate A_i's randomly, $b_i = \mathrm{tr}(A_i X^*)$
- MIT Logo
  - X^* obtained using MIT Logo image of size 38×73 and rank 4
  - Generate A_i's randomly, $b_i = \mathrm{tr}(A_i X^*)$

Figure: MIT Logo

- Compare against an adaptation of SVT (trace-norm relaxation)



(a)



(b)

Figure: **(a)**: Time taken by SVP and SVT for random instances optimal rank $k = 5$, **(b)**: Error for MIT logo

Results: Matrix Completion

- Synthetic Datasets:
  - Generate a random X^* of rank k
  - Generate Ω uniformly with sampling density p
- MovieLens Dataset:
  - User-movie ratings matrix
  - 1 million ratings for 3900 movies by 6040 users
- Compare against:
  - SVT (Cai et al.'08)
  - SMC (Keshavan et al.'09)
  - ADMiRA (Lee and Bresler'09)
  - Alternative least squares (ALS): our implementation

Results: Matrix Completion for Synthetic Datasets

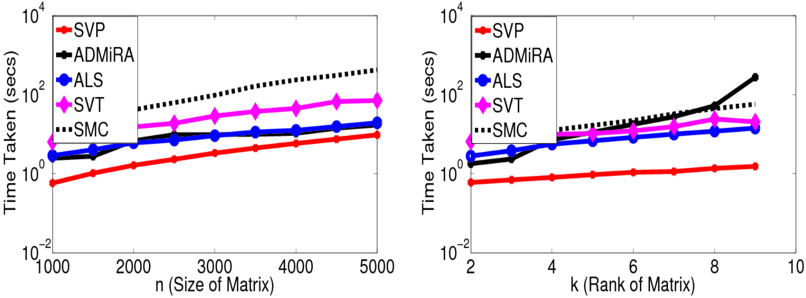


Figure: Running time (log scale) for different sizes and ranks

Results: Matrix Completion for Noisy Synthetic Datasets

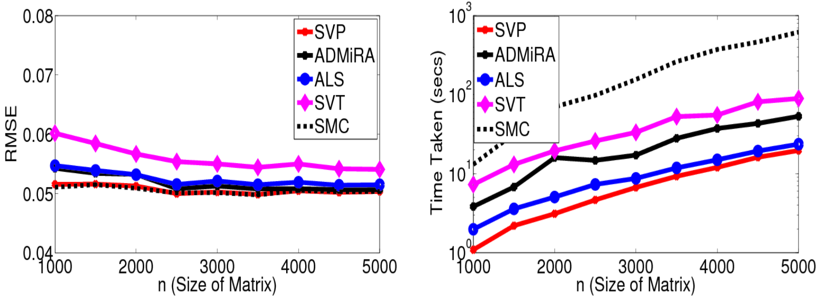


Figure: Noise level: 10% corrupt samples

Results: Matrix Completion for MovieLens Dataset

Method	RMSE	Time
SVP	1.01	64.85
SVT	1.21	1214.78
ALS	0.90	195.34

Table: RMSE obtained and Time taken by various methods

- Problem: Ratings matrix is not sampled uniformly

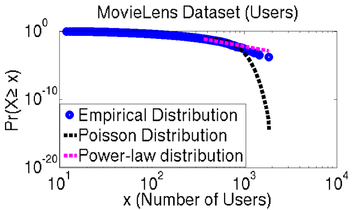


Figure: Cumulative degree distribution of users (MoviesLens)

Conclusions and Future Work

- Singular Value Projection (SVP) algorithm
  - Simple analysis for ARMP (with RIP)
  - Partial progress for matrix completion
  - SVP much faster than existing methods
- Future Work
  - Optimality of SVP for matrix completion
  - Other sampling distributions: power-law distributions
  - Hard thresholding algorithms for other problems, e.g., sparse+low-rank matrix decomposition

Paper available at: http://arxiv.org/abs/0909.5457
Code available at: http://www.cs.utexas.edu/users/pjain/svp/