Fast First-Order Methods for Stable Principal Component Pursuit

Donald Goldfarb
Joint work with Necdet Serhat Aybat and Garud Iyengar

Columbia University

Duke Workshop on
Sensing and Analysis of High-Dimensional Data
July 26-28, 2011

Overview

- Interested in fast algorithms for solving:

$$\min_{X,S\in\mathbb{R}^{m\times n}}\{\|X\|_* + \xi\|S\|_1 : \ \|X+S-D\|_F \leq \delta\},$$

for a given data matrix $D \in \mathbb{R}^{m\times n}$ and noise parameter δ.

- Why?
Many applications requires decomposing noisy D into low-rank and sparse components:

  -Video survellience

  -Face recognition

  -Ranking and collaborative filtering

Low Rank Matrix Completion

$$\min_{X \in \Re^{n \times n}} \left\{ \|X\|_* : X_{ij} = D_{ij}, (i,j) \in \Omega \right\}$$

- Unknown data matrix: $D \in \mathbb{R}^{n \times n}$, $\mathbf{rank}(D) = r \ll n$
- Observations: D_{ij} for all $(i,j) \in \Omega$
  $|\Omega| = \mathcal{O}(n^{1.2} r \log(n)) \ll n^2$.
- With high probability, unique optimal solution $X^* = D$
- Applications: Netflix problem, Sensor Localization

Robust Principal Component Analysis

Principal Component Pursuit:

$$\min_{X,S \in \Re^{n \times n}} \left\{ \|X\|_* + \xi \| \mathbf{vec}(S)\|_1 : \; X + S = D \right\}$$

- Data matrix: $D \in \mathbb{R}^{n \times n}$, $D = \bar{X} + \bar{S}$
- $\mathbf{rank}(\bar{X}) \ll n$, $\|\bar{S}\|_0 \ll n^2$
- With high probability, unique optimal solution $(X^*, S^*) = (\bar{X}, \bar{S})$
- Applications: Video surveillance, Ranking, Face recognition

Stable Principal Component Pursuit (SPCP)

$$\min_{X,S\in\Re^{n\times m}} \left\{ \|X\|_* + \xi\|S\|_1 : \ \|X+S-D\|_F \leq \delta \right\}$$

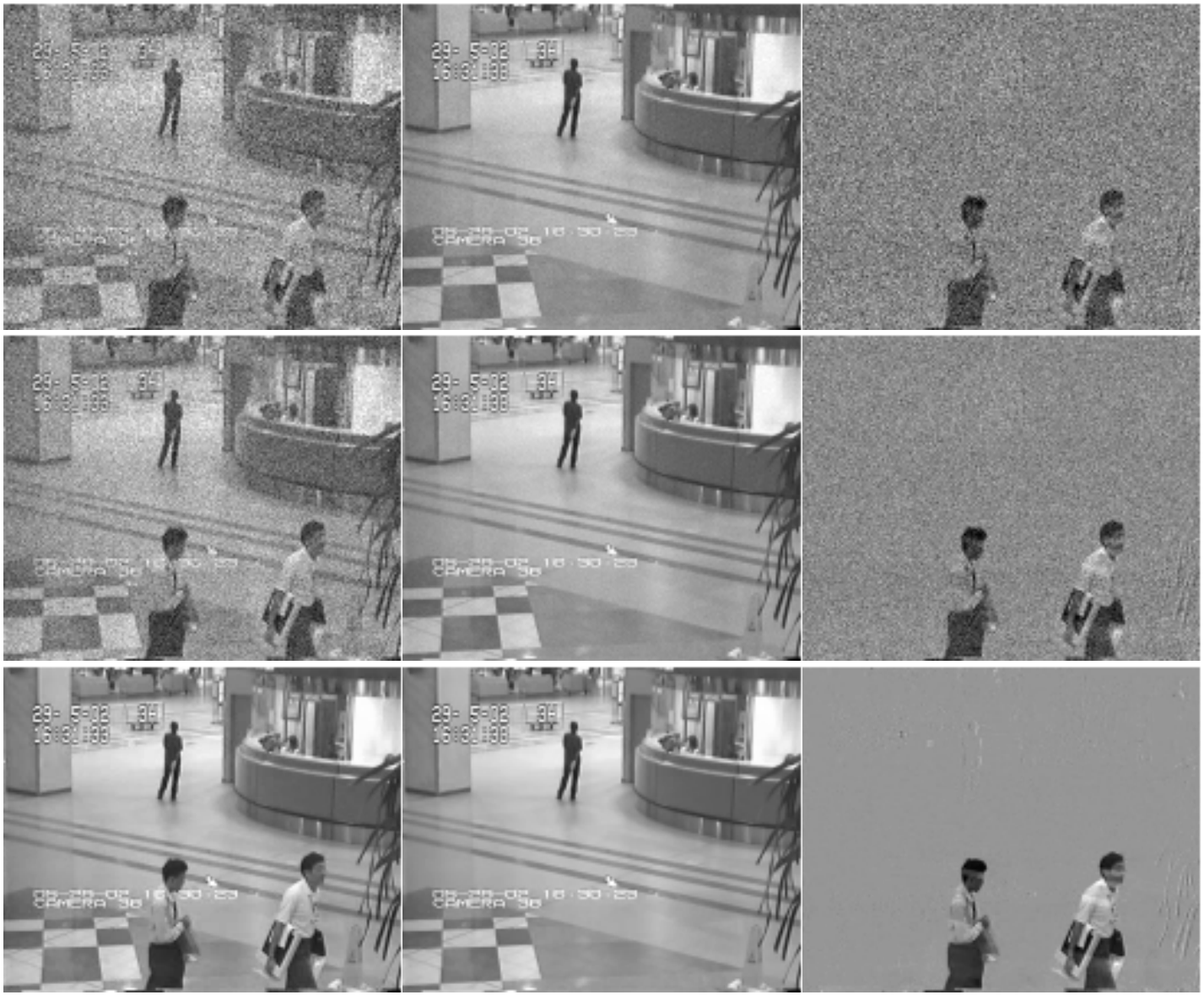
- Data matrix: $D \in \mathbb{R}^{n\times n}$, $D = \bar{X} + \bar{S} + \zeta$
- ζ i.i.d. noise matrix, $\|\zeta\|_F \leq \delta$
- $\mathbf{rank}(\bar{X}) \ll \min\{m,n\}$, $\|\bar{S}\|_0 \ll mn$
- With high probability, unique optimal solution (X^*, S^*) satisfies: $\|X^* - \bar{X}\|_F^2 + \|S^* - \bar{S}\|_F^2 \leq Cmn\delta^2$
- Applications: Video surveillance, Ranking, Face recognition

Videos with different levels of noise

15dB:

20dB:

∞dB:



SDP Formulation of SPCP

Stable Principal Component Pursuit is an SDP,

$$\begin{array}{ll} \min_{X,S\in\Re^{m\times n}} & \|X\|_* + \xi\|S\|_1 \\ s.t. & \|X+S-D\|_F \leq \delta \end{array}$$

$$\Updownarrow$$

$$\begin{array}{ll} \min_{X,S,W_1,W_2} & \frac{1}{2}\left(Tr(W_1)+Tr(W_2)\right) + \langle E, S_+ + S_- \rangle \\ s.t. & \|X+S_+-S_--D\|_F \leq \delta \\ & \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq \mathbf{0} \\ & S_+ \geq \mathbf{0},\ S_- \geq \mathbf{0} \end{array}$$

Previous Work

Only specialized algorithm: ASALM by Tao and Yuan.

- ASALM does alternating minimizations in X, S, Z directions on the augmented Lagrangian of

  $$\min_{X \in \mathbb{R}^{m \times n}} \{\|X\|_* + \xi \|S\|_1 : \ X + S + Z = D, \ \|Z\|_F \leq \delta\}$$

- ASALM iterates are not feasible
- ASALM converges to an optimal solution
- Complexity of ASALM is not known

Fact: There is **no** specialized algorithm for SPCP with a known iteration complexity bound.

Question: Can one achieve **work/iteration** $\propto$ one gradient computation with existing first-order algorithms that have **low iteration complexities**?

Nesterov's Fast Gradient Algorithms

Problem: $\min_{x \in Q} p(x) + f(x)$, where p, f are closed convex functions, Q is a closed convex set and ∇f is L-Lipschitz continuous.

PROXIMAL GRADIENT ALGORITHM (x_0)

1: **while** $(k \geq 0)$ **do**
2: $\quad y_k \leftarrow \operatorname{argmin}_{x \in Q} p(x) + f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|_2^2$
3: $\quad z_k \leftarrow \operatorname{argmin}_{x \in Q} \frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^{k} \frac{i+1}{2}[p(x) + f(x_i) + \langle \nabla f(x_i), x - x_i \rangle]$
4: $\quad x_{k+1} \leftarrow \frac{2}{k+3} z_k + \frac{k+1}{k+3} y_k$
5: $\quad k \leftarrow k + 1$
6: **end while**

Let $x^* = \operatorname{argmin}_{x \in Q}\{p(x) + f(x)\}$. Then for all $k \geq \sqrt{\frac{2L\|x_0 - x^*\|_2^2}{\epsilon}}$, we have $p(y_k) + f(y_k) \leq p(x^*) + f(x^*) + \epsilon$.
(**Analogous result for FISTA**)

Per iteration complexity depends on the complexity of projection onto Q and computing $\nabla f(x_k)$.

First-Order Algorithms

$$f_\mu(X) := \max\{\langle X, U\rangle - \tfrac{\mu}{2}\|U\|_F^2 : \ \|U\|_2 \le 1\},$$
$$g_\nu(S) := \max\{\langle S, W\rangle - \tfrac{\nu}{2}\|W\|_F^2 : \ \|W\|_\infty \le 1\}.$$

- **Nesterov's algorithm**:
  $\min_{X,S}\{f_\mu(X) + \xi g_\nu(S) : \ (X,S) \in \chi\}$
  If $\mu = \nu = \Omega(\epsilon)$, then ϵ-optimal in $\mathcal{O}(1/\epsilon)$ iterations
- **FISTA**:
  $\min_{X,S}\{f_\mu(X) + \xi\|S\|_1 : \ (X,S) \in \chi\}$
  If $\mu = \Omega(\epsilon)$, then ϵ-optimal in $\mathcal{O}(1/\epsilon)$ iterations
- **FALM-S algorithm with partial splitting**:
  $\min_{X,Z,S}\{f_\mu(X) + \xi\|S\|_1 : \ X = Z, \ (Z,S) \in \chi\}$
  If $\mu = \Omega(\epsilon)$, then ϵ-optimal in $\mathcal{O}(1/\epsilon)$ iterations

Complexity of solving subproblems?

Efficient Solution of Subproblems

Key Lemma:
$\chi = \{(X, S) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} : \ \|X + S - D\|_F \leq \sigma\}.$

$$(I) \quad : \quad \min \left\{ \|S - \tilde{S}\|_F^2 + \|X - \tilde{X}\|_F^2 : \ (X, S) \in \chi \right\}.$$
$$(II) \quad : \quad \min \left\{ \xi\|S\|_1 + \tfrac{\rho}{2}\|X - \tilde{X}\|_F^2 : \ (X, S) \in \chi \right\}.$$

Optimal solutions of (I) and (II) can be computed in $\mathcal{O}(\mathbf{mn})$ and $\mathcal{O}(\mathbf{mn}\log(\mathbf{mn}))$, respectively.

Solutions if (I) or (II) are needed depending on whether Nesterov, FISTA or FALM-S is used.

Euclidean Projection onto χ in $\mathcal{O}(\mathbf{mn})$:

Let $(X^*, S^*) = \operatorname{argmin}_{X,S} \left\{ \|S - \tilde{S}\|_F^2 + \|X - \tilde{X}\|_F^2 : \ (X, S) \in \chi \right\}$

When $\delta > 0$,

$$X^* = \left(\frac{\theta^*}{1 + 2\theta^*} \right) \left(D - \tilde{S} \right) + \left(\frac{1 + \theta^*}{1 + 2\theta^*} \right) \tilde{X},$$

$$S^* = \left(\frac{\theta^*}{1 + 2\theta^*} \right) \left(D - \tilde{X} \right) + \left(\frac{1 + \theta^*}{1 + 2\theta^*} \right) \tilde{S},$$

$$\theta^* = \max \left\{ 0, \ \frac{1}{2} \left(\frac{\|\tilde{X} + \tilde{S} - D\|_F}{\delta} - 1 \right) \right\}.$$

When $\delta = 0$,

$$X^* = \tfrac{1}{2} \left(D - \tilde{S} \right) + \tfrac{1}{2} \tilde{X} \text{ and } \quad S^* = \tfrac{1}{2} \left(D - \tilde{X} \right) + \tfrac{1}{2} \tilde{S}.$$

ℓ_1-Euclidean Projection onto χ in $\mathcal{O}(\mathbf{mn}\log(\mathbf{mn}))$:

Let $(X^*, S^*) = \operatorname{argmin}_{X,S} \left\{ \xi \|S\|_1 + \frac{\rho}{2}\|X - \tilde{X}\|_F^2 : \ (X, S) \in \chi \right\}$

When $\delta > \mathbf{0}$,

$$S^* = sign\left(D - \tilde{X}\right) \odot \max\left\{ |D - \tilde{X}| - \xi \frac{(\rho + \theta^*)}{\rho\theta^*} \ E, \ \mathbf{0} \right\},$$

$$X^* = \frac{\theta^*}{\rho + \theta^*} \ (D - S^*) + \frac{\rho}{\rho + \theta^*} \ \tilde{X},$$

$$\theta^* = \begin{cases} 0, & \|D - \tilde{X}\|_F \leq \delta; \\ \phi^{-1}(\delta), & \text{otherwise.} \end{cases}$$

where $\phi(\theta) := \| \min\left\{ \frac{\xi}{\theta} \ E, \ \frac{\rho}{\rho+\theta} \ |D - \tilde{X}| \right\} \|_F$ and $\mathbf{dom}(\phi) = \mathbb{R}_{++}$.

When $\delta = \mathbf{0}$,

$$S^* = sign\left(D - q(\tilde{X})\right) \odot \max\left\{ |D - q(\tilde{X})| - \frac{\xi}{\rho} \ E, \ \mathbf{0} \right\},$$

$$X^* = D - S^*.$$

Non-Smooth Augmented Lagrangian (NSA) algorithm

Split X and apply alt. direction augmented Lagrangian method to

Equivalent problem:

$\min_{X,S,Z}\{\|X\|_* + \xi\|S\|_1 : \ X = Z, \ (Z,S) \in \chi\}$

NSA Algorithm (X_0, S_0)

1: **while** $(k \geq 0)$ **do**
2: $\quad X_{k+1} \leftarrow \min_X\{\|X\|_* + \langle Y_k, X - Z_k\rangle + \frac{\rho_k}{2}\|X - Z_k\|_F^2\}$
3: $\quad (Z_{k+1}, S_{k+1}) \leftarrow$
$\quad \operatorname{argmin}_{Z,S}\{\xi\|S\|_1 + \langle Y_k, X_{k+1} - Z\rangle + \frac{\rho_k}{2}\|X_{k+1} - Z\|_F^2 : \ (Z,S) \in \chi\}$
4: $\quad Y_{k+1} \leftarrow Y_k + \rho_k(X_{k+1} - Z_{k+1})$
5: $\quad$ Choose ρ_{k+1} such that $\rho_{k+1} \geq \rho_k$
6: $\quad k \leftarrow k + 1$
7: **end while**

NSA convergence result

Theorem: Let $\{X_k, Z_k, S_k, Y_k\}_{k\in\mathbb{Z}_+}$ be the sequence produced by Algorithm NSA and let (X^*, S^*) be a solution to SPCP.

(i) If $\sum_{\mathbf{k}\in\mathbb{Z}_+} \frac{\mathbf{1}}{\rho_{\mathbf{k}}} = \infty$, then $X_k \to X^*$, $Z_k \to X^*$, $S_k \to S^*$.

(ii) If $\sum_{\mathbf{k}\in\mathbb{Z}_+} \frac{\mathbf{1}}{\rho_{\mathbf{k}}^{\mathbf{2}}} = \infty$ **and** $\|\mathbf{D} - \mathbf{X}^*\|_{\mathbf{F}} \neq \delta$, then $Y_k \to Y^*$ (optimal Lagrangian multiplier).

Experimental Setup

Data Matrix: $D = \bar{X} + \bar{S} + \zeta$,

(i) $\bar{X} = UV^T$, where $U \in \Re^{n \times r}$, $V \in \Re^{n \times r}$
$U_{ij} \sim N(0,1)$, $V_{ij} \sim N(0,1)$ for all i, j

(ii) $\Lambda \subset \{(i,j) : i,j = 1, \ldots, n\}$, $|\Lambda| = p$ chosen randomly

(iii) $\bar{S}_{ij} \sim U[-100, 100]$ for all $(i,j) \in \Lambda$

(iv) $\zeta_{ij} \sim \sigma N(0,1)$ for all i, j

Create 10 random $D \in \Re^{n \times n}$ s.t. $r = c_r n$, $p = c_p n^2$.

- $n \in \{500, 1000, 1500, 2000\}$
- $c_r \in \{0.05,\ 0.1\}$
- $c_p \in \{0.05,\ 0.1\}$
- $\sigma = 10^{-3}$

Numerical Results for NSA

Table: Average # svd/cpu(sec) for decomposing $D = \bar{X} + \bar{S} + \zeta$

	$c_r = 0.05$		$c_r = 0.1$	
	$c_p = 0.05$	$c_p = 0.1$	$c_p = 0.05$	$c_p = 0.1$
n	**# svd /cpu**	**# svd /cpu**	**# svd/cpu**	**# svd/cpu**
500	11/5.7	11.9/6.4	12.2/6.5	13/6.9
1000	11.8/21.7	12.7/24	13/31.4	14.1/36.1
1500	12.8/54.6	12.9/52.2	14/95.1	15/100.4
2000	12.9/115.7	13/114.3	14/206.6	15/223.9

The solution accuracy:
$\frac{\|X^{sol} - \bar{X}\|_F}{\|\bar{X}\|_F} = 5 \times 10^{-5}$, $\frac{\|S^{sol} - \bar{S}\|_F}{\|\bar{S}\|_F} = 2 \times 10^{-5}$

Figure: $D \in \mathbb{R}^{n \times n}$, $n = 1500$, $\sigma = 1 \times 10^{-3}$, $SNR \approx 80dB$

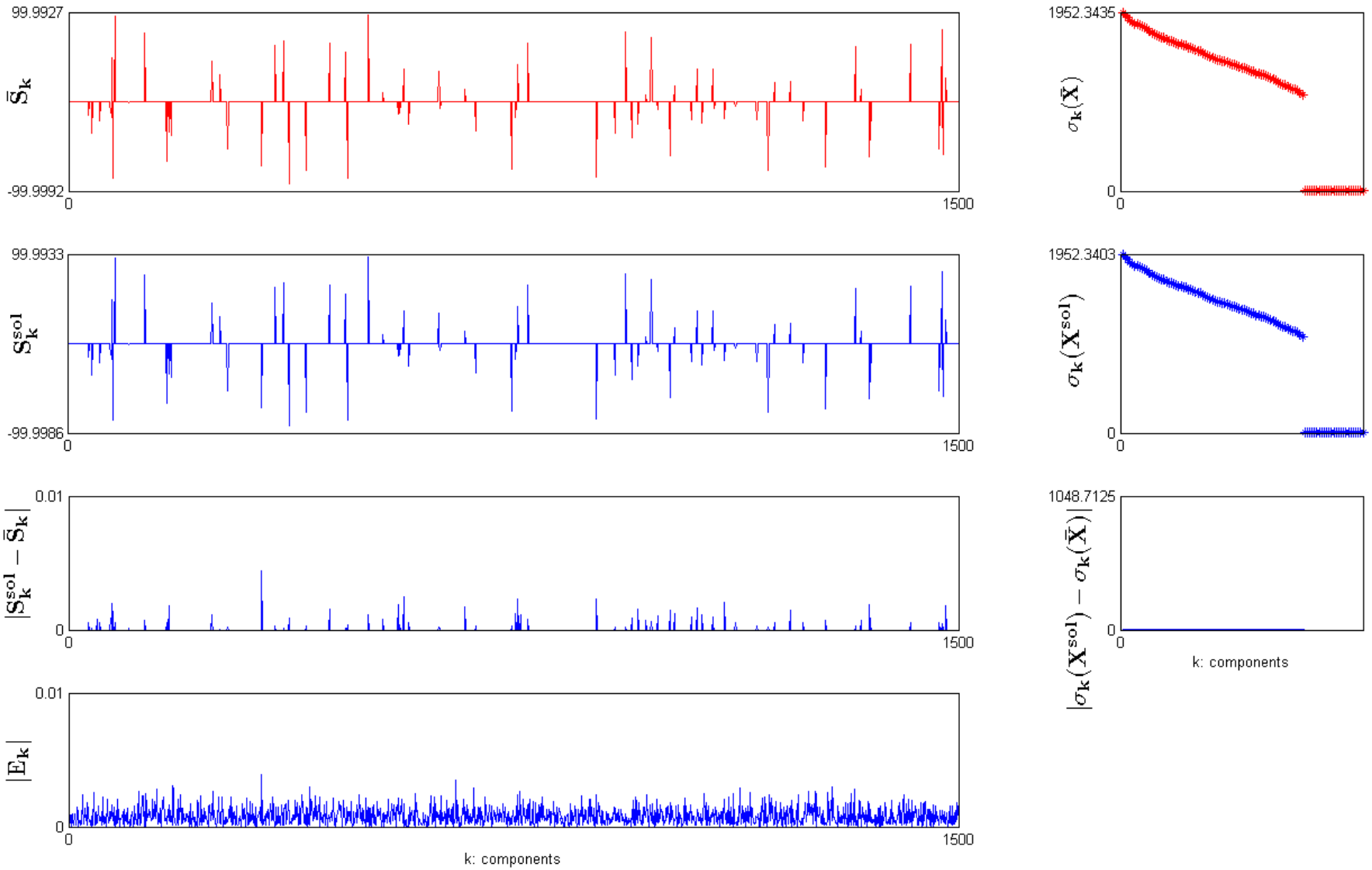
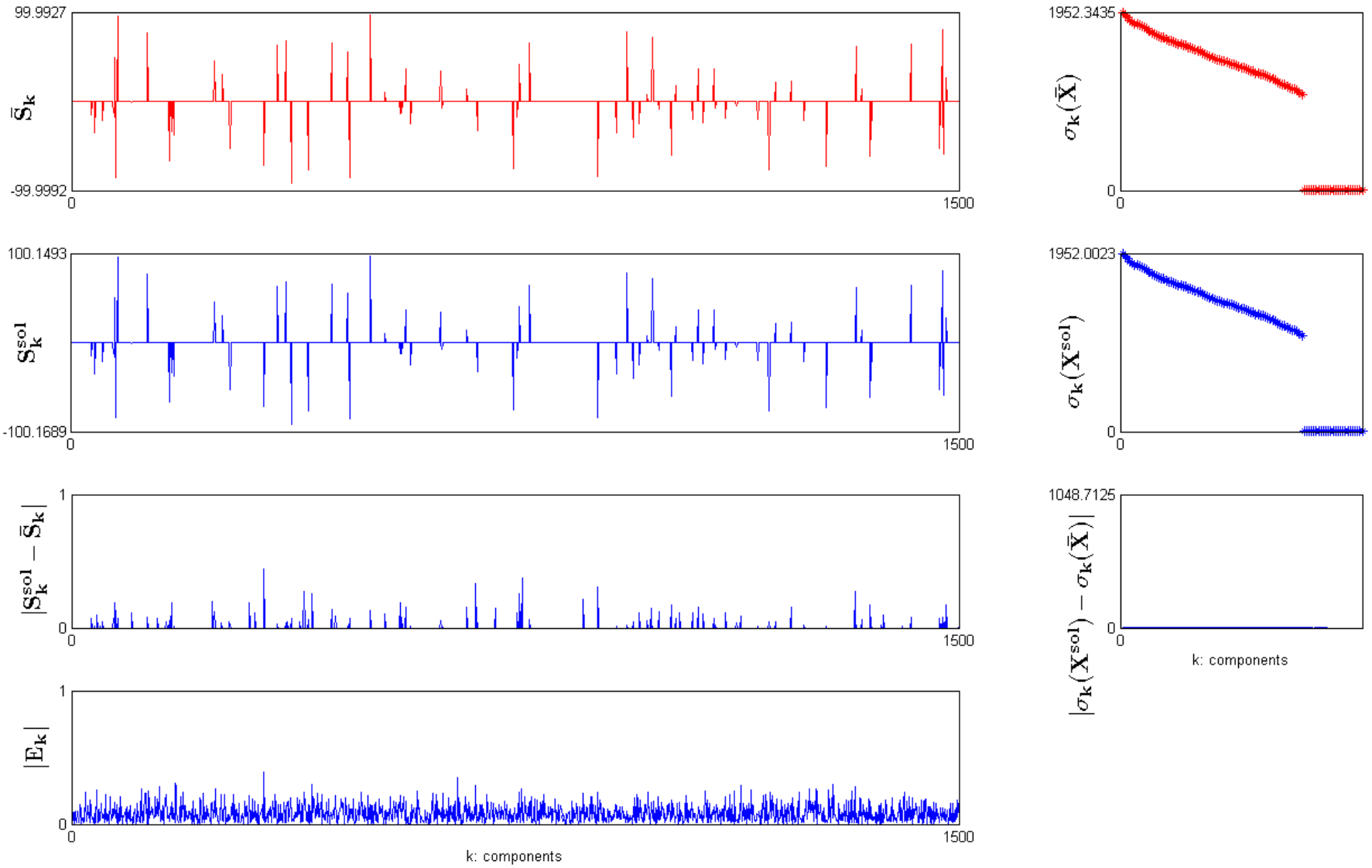
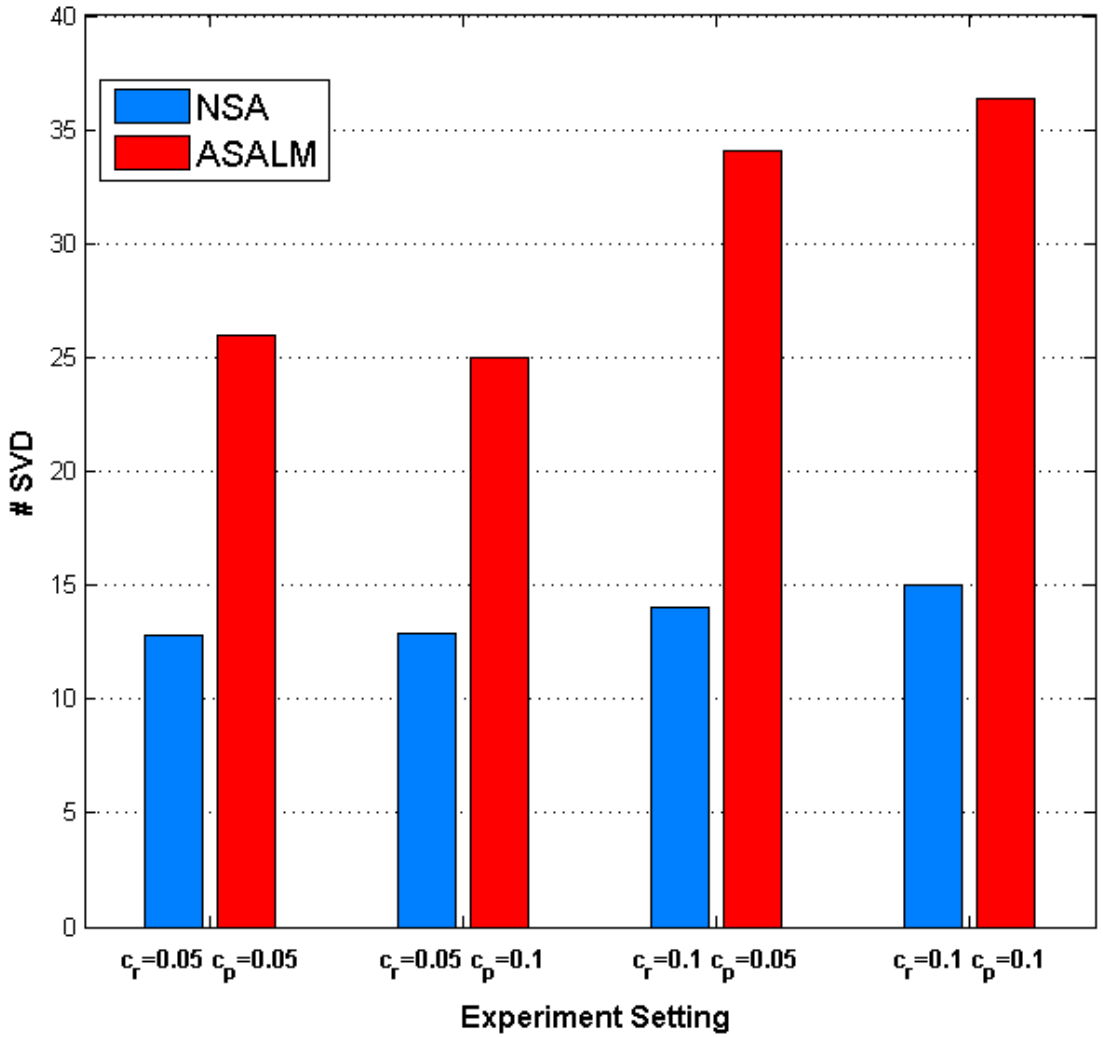


Figure: $D \in \mathbb{R}^{n \times n}$, $n = 1500$, $\sigma = 1 \times 10^{-1}$, $SNR \approx 40dB$



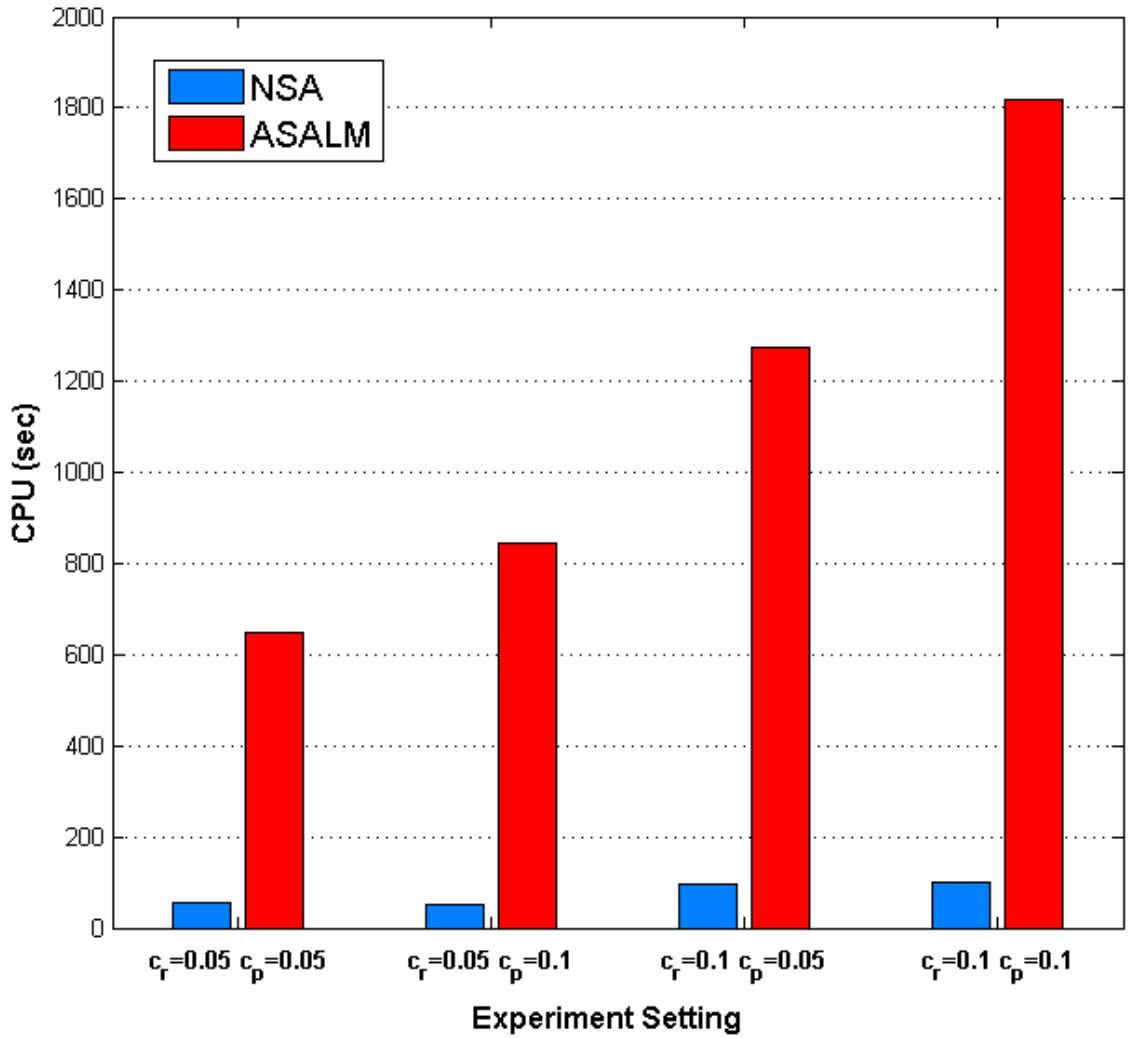
Numerical Results: NSA vs ASALM

Average # svd for decomposing $D = \bar{X} + \bar{S} + \zeta$, $n = 1500$, $\sigma = 1 \times 10^{-3}$, $SNR \approx 80dB$



Numerical Results: NSA vs ASALM

Average CPU (sec) for decomposing $D = \bar{X} + \bar{S} + \zeta$, $n = 1500$, $\sigma = 1 \times 10^{-3}$, $SNR \approx 80dB$



Video Surveillance Example

- T: number of frames
- $N \equiv m \times n$ is the frame resolution

To detect moving objects

i. Form i-th column of $D \in \mathbb{R}^{N \times T}$ by stacking the columns of i-th frame.

ii. Solve $\min\{\|X\|_* + \frac{1}{\sqrt{\max\{N,T\}}}\|S\|_1 : \ \|X + S - D\|_F \leq \delta\}$

Suppose there is no noise, i.e. $\delta = 0$, $\bar{X} + \bar{S} = D$. Then

- i-th column of $\bar{S}$ is the moving object in the i-th frame
- i-th column of $\bar{X}$ is the background in the i-th frame

Note that $\bar{X}$ is a **low-rank** matrix.

Noiseless Video

D(t):

X(t):

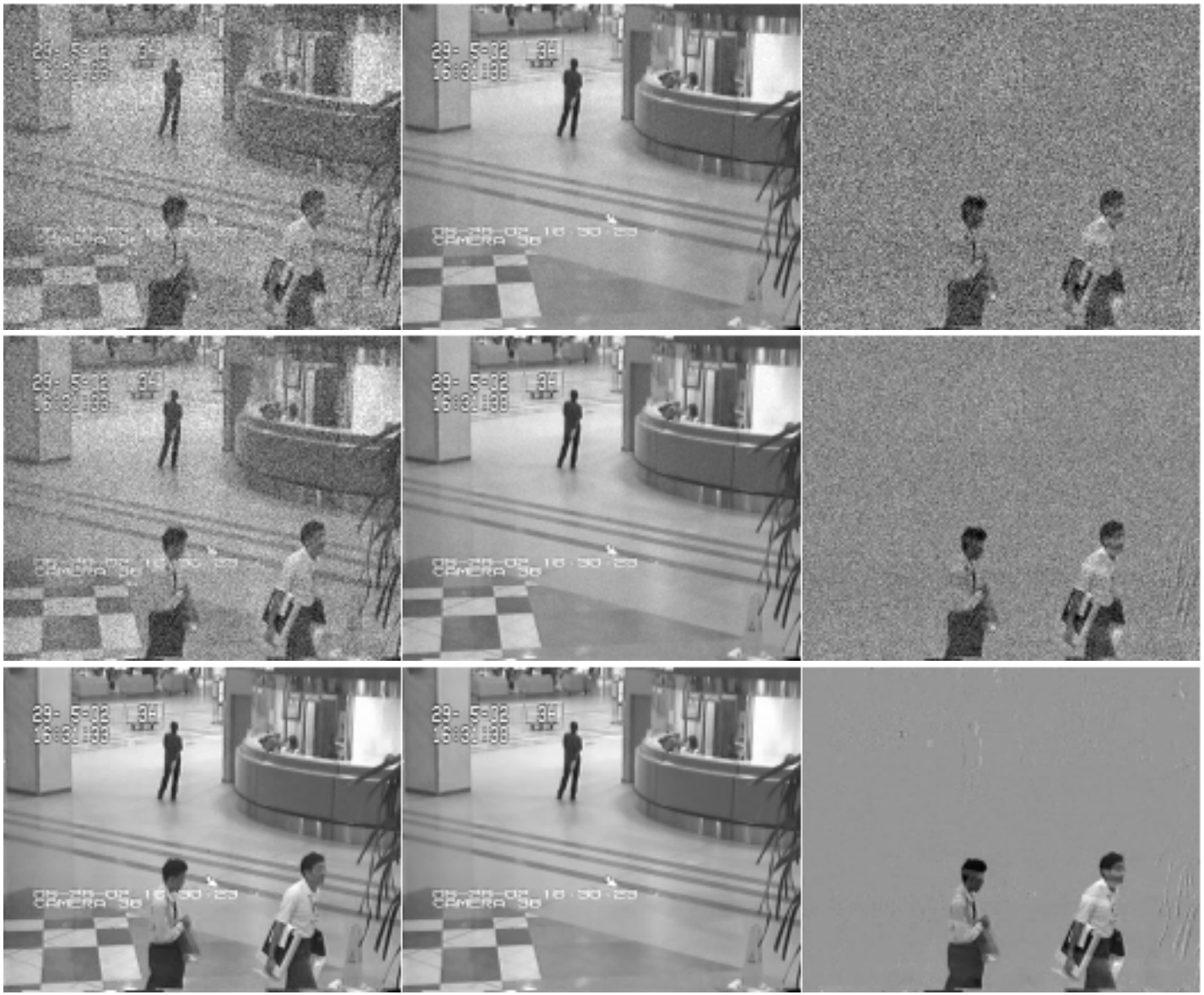
S(t):

Videos with Different Noise Levels

15dB:

20dB:

∞dB:



References

- N. S. Aybat, D. Goldfarb, G. Iyengar, Fast First-Order Methods for Stable Principal Component Pursuit `http://arxiv.org/abs/1105.2126`

Thank You