RGNMR: A Provable Algorithm for Robust Matrix Completion

Eilon Vaknin Laufer   Boaz Nadler
Weizmann Institute of Science

Robust Matrix Completion (RMC)

Let $X = L^* + S^* \in \mathbb{R}^{n_1 \times n_2}$ where L^* is of rank r and S^* is a corruption matrix with a few non-zero entries at unknown locations. Let $\Omega \subset [n_1] \times [n_2]$ be a subset of observed entries.

Problem: Recover L^* from a subset of observed entries $\{X_{i,j} \mid (i,j) \in \Omega\}$.

Applications

Recovering a low rank matrix from a subset of its entries has applications in recommendation systems, various problems in computer vision and sensor network localization. A key challenge in these and other applications is that some of the observed entries may be arbitrarily corrupted outliers.

Previous Algorithms

`AOP` [Yan, Yang, and Osher 2013], `RPCA-GD` [Yi et al. 2016], `RMC` [Cambier and Absil 2016], `RRMC`[Cherapanamjeri, Gupta, and Jain 2017], `HUB`[Ruppel, Muma, and Zoubir 2020], `HOAT` [Z.-Y. Wang, Li, and So 2023] and others.

Limitations of Existing Methods

- Require large number of observed entries.
- Require the (often unknown) rank r of L^*, fail when overparameterized even with an input rank of $r+1$.
- Fail to recover the matrix L^* if it has a moderate condition number, as low as 5.

Our Contributions

- Propose `RGNMR`, a new RMC method that overcomes the above limitations.
- Developed a scheme to estimate the number of corrupted entries in X
- Derived recovery guarantees for `RGNMR` which improve upon the best currently known for other (factorization-based) methods.

RGNMR

Working variables: L an estimate of L^* and $\Lambda \subset \Omega$, an estimate of the locations of corrupted entries.

`RGNMR` **iterates these two steps :**

- Step I : Given the current set of suspected outlier entries Λ, update L using the remaining entries $\{X_{i,j} \mid (i,j) \in \Omega \setminus \Lambda\}$.
- Step II: given the updated matrix L, recompute the set of suspected outliers Λ, by the k entries with largest magnitude in $\{(L-X)_{i,j} \mid (i,j) \in \Omega\}$

Algorithm - `RGNMR`

Input:

- $\{X_{i,j} \mid (i,j) \in \Omega\}$ - observed entries
- r - rank of L^*
- k - assumed number of corrupted entries
- $\begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \in \mathbb{R}^{n_1+n_2 \times r}$ factor matrices of initial estimate of L^*.
- Λ_0 - initial estimate of the set of corrupted entries
- T - maximal number of iterations

Output: $\hat{L}$ of rank r

for $t = 0 \ldots T-1$ **do**

$$\begin{pmatrix} U_{t+1} \\ V_{t+1} \end{pmatrix} = \arg\min_{U,V} \|U_t V^\top + U V_t^\top - U_t V_t^\top - X\|^2_{F(\Omega \setminus \Lambda_t)}$$

$$\Lambda_{t+1} = \arg\min_{\Lambda \subset \Omega, |\Lambda| = k} \|U_t V_{t+1}^\top + U_{t+1} V_t^\top - U_t V_t^\top - X\|^2_{F(\Omega \setminus \Lambda)}$$

end for

return $P_r(U_{T-1} V_T^\top + U_T V_{T-1}^\top - U_{T-1} V_{T-1}^\top)$

Estimating the Number of Corrupted Entries

Problem: Finding a tight upper bound on the true number of corrupted entries k^*.

Observation: Empirically, the estimates Λ_t of the set of corrupted entries converge if and only if $k \le k^*$.

Our Solution: Binary search for k^*.

Formally, set $k_{\min} = 0$ and $k_{\max} = |\Omega|/2$. Run `RGNMR` with $k = \lfloor (k_{\min} + k_{\max})/2 \rfloor$. If Λ_t converged, update $k_{\min} = k$. Otherwise, set $k_{\max} = k$. Run till convergence.

Assumptions for Theoretical Analysis

- The underlying matrix L^* has an incoherence parameter μ.
- [Bernoulli Model] Each entry of X is independently observed with probability p.
- In each row and column, the fraction of observed entries which are corrupted is bounded by $\alpha \in (0,1)$.

We denote by $\mathcal{M}(n_1, n_2, r, \mu, \kappa)$ the set of $n_1 \times n_2$ matrices of rank r, incoherence parameter μ and condition number κ.

Theorem (Informal)

Let $X = L^* + S^*$, where $L^* \in \mathcal{M}(n_1, n_2, r, \mu, \kappa)$, $n_1 \ge n_2$. There exist constants C, c_α such that : If the fraction of corrupted entries is small enough, $\alpha < \frac{1}{c_\alpha r \mu \kappa}$ and the probability to observe an entry is high enough, $p \ge \frac{C\mu r}{n_2} \max\{\log n_1, \mu r \kappa^2\}$, then w.p. at least $1 - \frac{6}{n_1}$, `RGNMR` with suitable initialization converges linearly to L^*.

Simulations Results

L^* is a rank 5 matrix of size 3200×400. The fraction of corrupted entries is $\alpha = 5\%$. The oversampling ratio is $\frac{|\Omega|}{r \cdot (n_1 + n_2 - r)} = 12$ and the condition number is $\kappa = 2$. For any RMC method which outputs $\hat{L}$, we compute two performance measures:

(i) Median relative reconstruction error $\texttt{rel-RMSE} = \frac{\|\hat{L} - L^*\|_F}{\|L^*\|_F}$;

(ii) Failure probability, $\mathbb{P}(\texttt{rel-RMSE} > 10^{-3})$, error bars of 95% confidence interval.

`RGNMR` is given the true number of corrupted entries k^* while `RGNMR-BS` is given an estimate of k^* obtained by our binary search scheme.

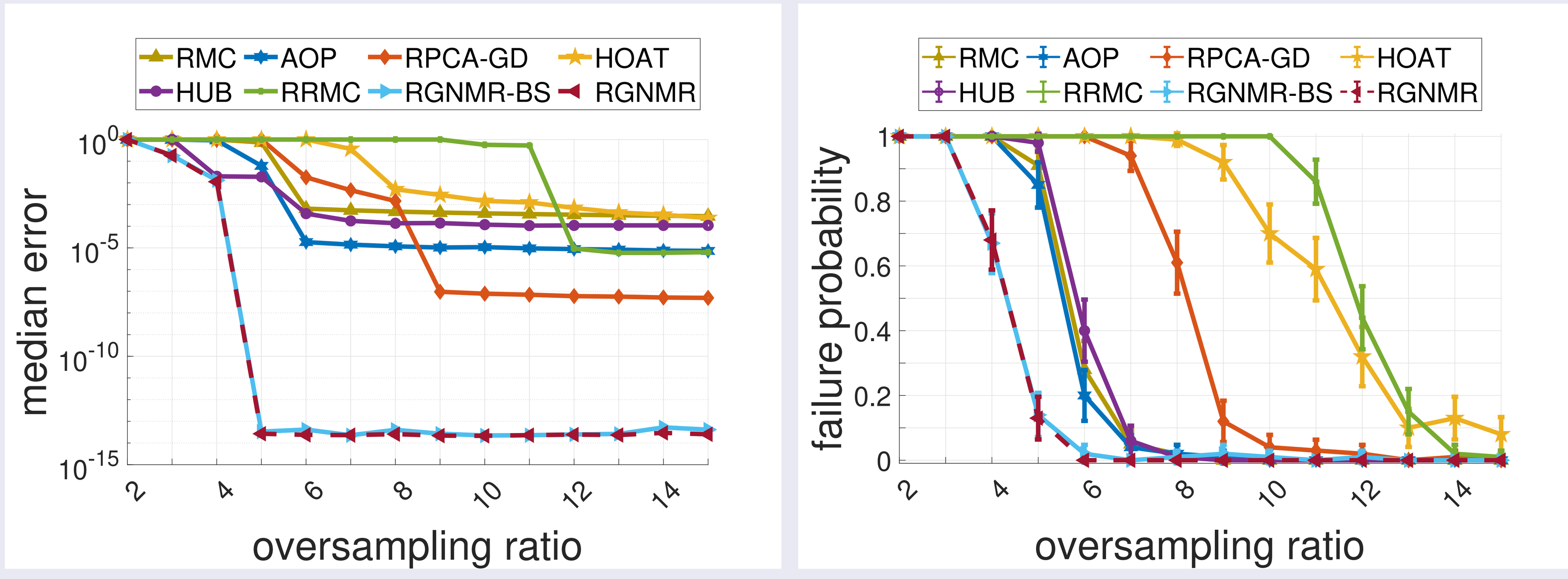


Figure: Performance as a function of the oversampling ratio $\frac{|\Omega|}{r \cdot (n_1 + n_2 - r)}$.

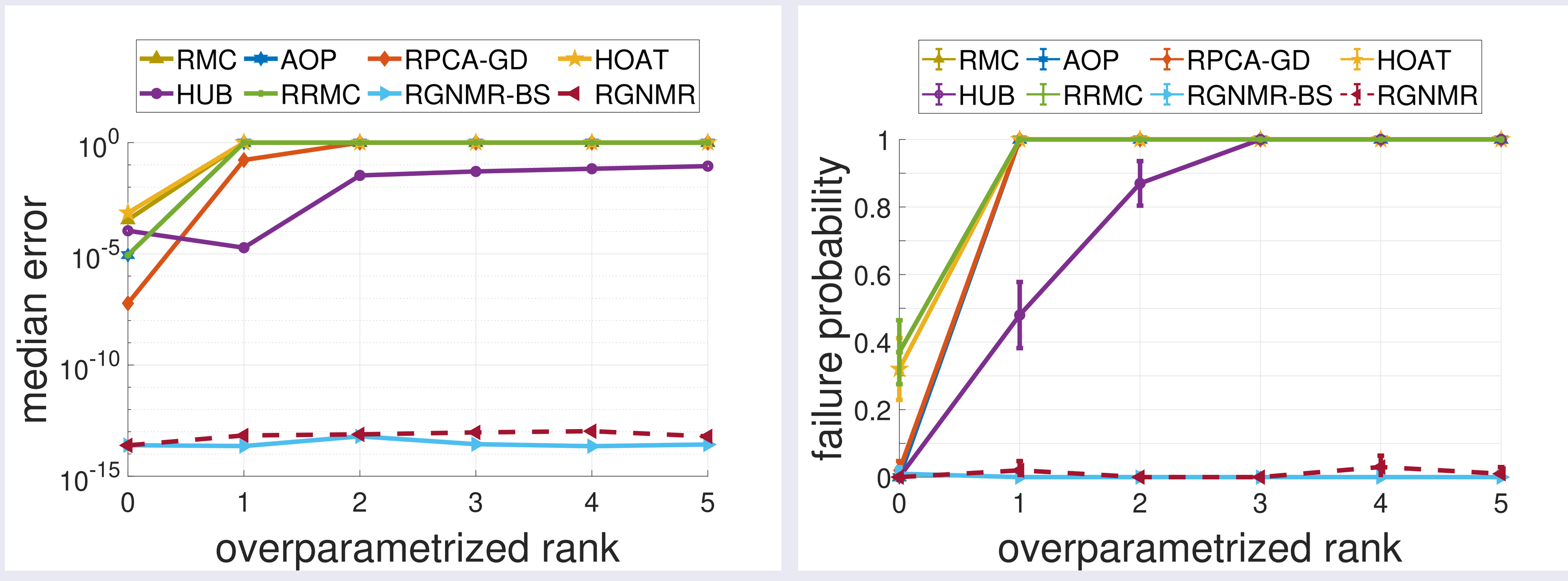


Figure: Performance under overparameterization. The input rank is $5 + i$ for $i \in [0, 5]$.

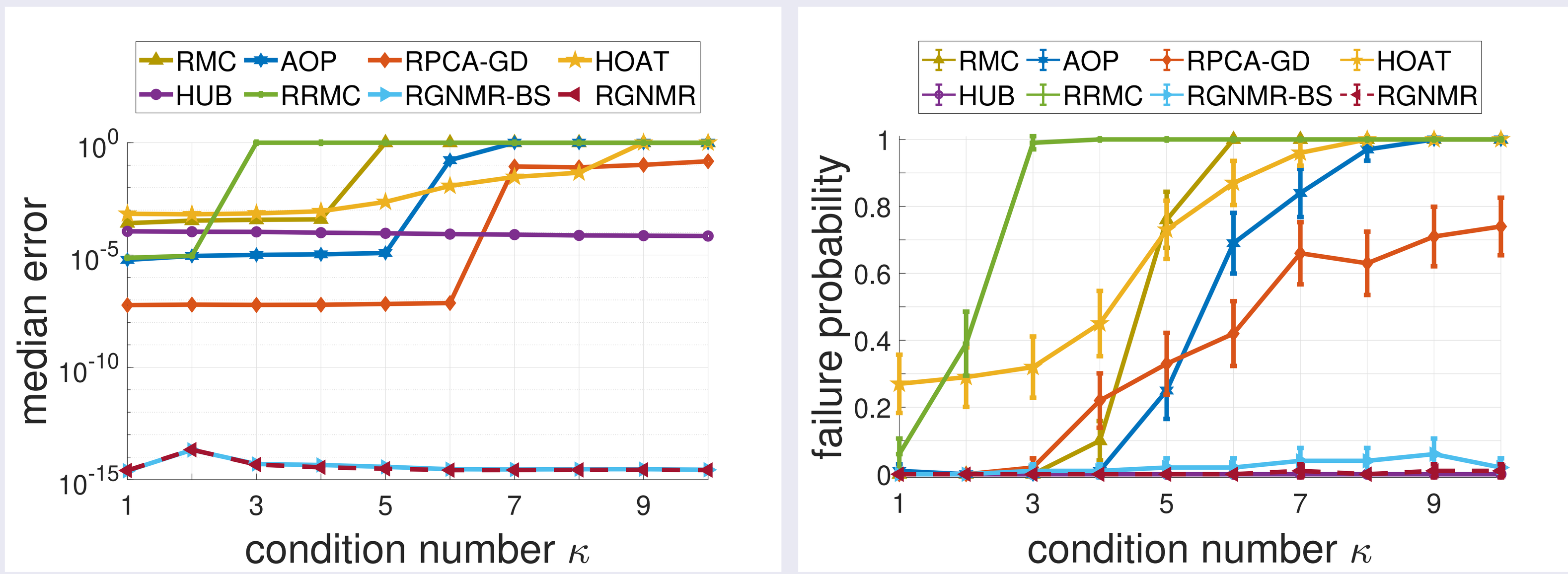
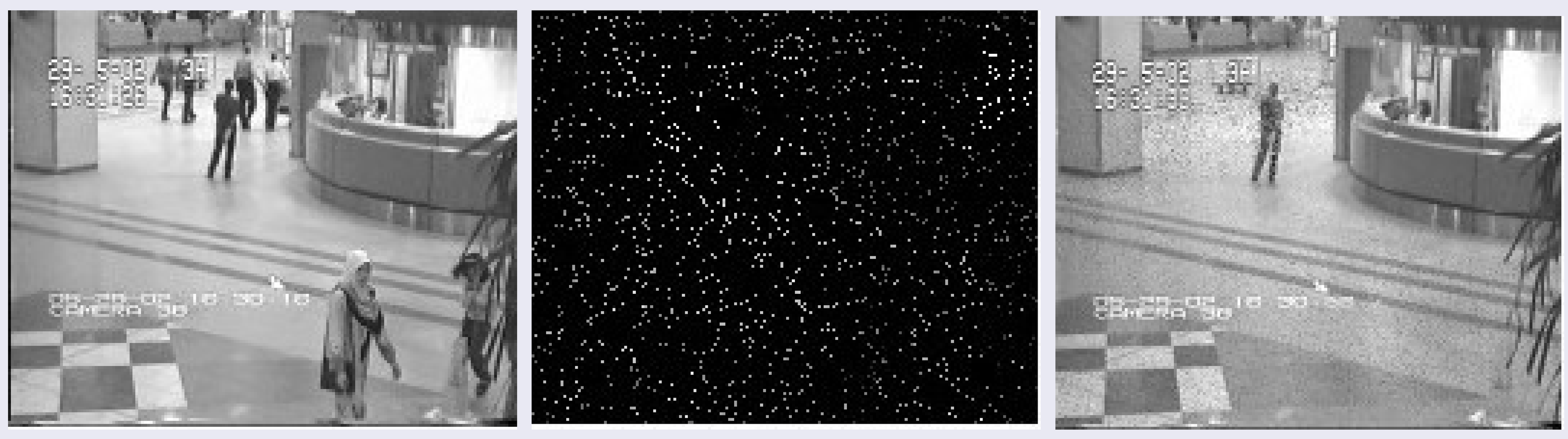


Figure: Performance as a function of the condition number κ.

Additionally `RGNMR` can handle a large fraction of corrupted entries, non uniform sampling, additive noise and high rank matrices.

Background Extraction

`RGNMR` also performs well on a real dataset involving background extraction in a video. This is a standard benchmark for RMC methods. The frames in the video can be decomposed to a low rank matrix corresponding to the static background plus a sparse matrix corresponding to the moving foreground.



(a) Original Image   (b) Sampled Image   (c) `RGNMR`

Figure: Background extraction for "Hall" video data. The frames are recovered from 5% of the original entries with an input rank of $r = 1$.

Comparison to Other Recovery Guarantees

Method Type	Method	Sample Complexity ($pn_2 \ge$)	Corruption Rate ($\alpha \le$)
Factorization Based	Zheng and Lafferty 2016	$\boldsymbol{\max\{\mu r \log n_1, \mu^2 r^2 \kappa^2\}}$	No Corruption
	Tong, Ma, and Chi 2021	Fully Observed	$\frac{1}{r^{\frac{3}{2}} \mu \kappa}$
	Cai et al. 2024	Fully Observed	$\frac{1}{r^{\frac{3}{2}} \mu \kappa}$
	Yi et al. 2016	$\mu^2 r^2 \kappa^4 \log n_1$	$\frac{1}{r \mu \kappa^2}$
	`RGNMR`	$\boldsymbol{\max\{\mu r \log n_1, \mu^2 r^2 \kappa^2\}}$	$\boldsymbol{\frac{1}{r \mu \kappa}}$
Full Matrix	Cherapanamjeri, Gupta, and Jain 2017	$\boldsymbol{\mu^2 r^2 \log^2(\mu r \sigma_1^*) \log^2 n_1}$	$\boldsymbol{\frac{1}{r \mu}}$
	T. Wang and Wei 2024	$\mu^3 r^3 \kappa^4 \log n_1$	$\frac{1}{r^2 \mu^2 \kappa^2}$

Table: Recovery guarantees requirements up to constant factors. Weakest conditions in each type of methods are in red bold.