Robust Outlier Identification for Noisy Data via Randomized Adaptive Compressive Sampling

Xingguo Li and Jarvis Haupt
Department of Electrical and Computer Engineering, University of Minnesota Twin Cities

***Abstract*—This paper examines the problem of locating outlier columns in a large, otherwise low-rank matrix, in the setting where the data are noisy. We propose a randomized two-step inference framework, and establish sufficient conditions on the required sample complexities under which these methods succeed (with high probability) in accurately locating the outliers. Numerical experimental results are provided to verify the theoretical bounds and demonstrate the computational efficiency of the proposed algorithm.**

I. Introduction

In this paper we examine a robust outlier identification problem. Given a data matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, we assume that

$$\mathbf{M} = \mathbf{L} + \mathbf{C} + \mathbf{N}, \tag{1}$$

where $\mathbf{L}$ is a rank-r matrix, $\mathbf{C}$ is a column-sparse matrix with k nonzero columns that are interpreted as "outliers" of the subspace spanned by columns of $\mathbf{L}$, and $\mathbf{N}$ is an additive noise. Our goal is to identify the locations $\mathcal{I}_{\mathbf{C}}$ of the nonzero columns of $\mathbf{C}$, *without* necessarily identifying the inliers (or the subspace they span), and n_1, n_2 are possibly very large relative to r and k.

Our investigation is motivated by a wide class of "big data" applications where the outliers themselves are of interest, such as collaborative filtering [1], network traffic [2], and computer vision [3], [4]. A number of contemporary methods have been developed, which exploit low-dimensional models within the context of convex inference methods based on robust PCA [5]–[8]. Despite their provable analytical successes, these methods can be computationally demanding when applied to very large data matrices.

Based on our initial investigation for noiseless case [9], [10], we propose a randomized two-step inference procedure having both low sample and implementation complexities, called robust adaptive compressive outlier sensing (RACOS), for locating column outliers from noisy observations. In Step 1, we compress the rows and sample a few columns of $\mathbf{M}$, and perform outlier pursuit (OP) to estimate the column space of the corresponding low-rank component. In Step 2, we perform a column-wise projection of the row compressed data of $\mathbf{M}$ onto the estimated column space obtained from Step 1 to recover the identities of outliers. The details are provided in Algorithm 1.

Algorithm 1 RACOS for Noisy Observations (RACOS-N)

Input: $\mathbf{M}$, $\gamma \in (0,1)$, $\lambda, \alpha, \varepsilon_1, \varepsilon_2 > 0$, and $q, m \in [n_1]$
Initialize: $\mathbf{\Phi} \in \mathbb{R}^{m \times n_1}$, $\mathbf{\Psi} \in \mathbb{R}^{q \times m}$ and $\mathbf{S} = \mathbf{I}_{:,\mathcal{S}}$, where $\mathcal{S} = \{j \in [n_2] : S_j \overset{iid}{\sim} \text{Bernoulli}(\gamma) = 1\}$ and $p = |\mathcal{S}|$
Step 1. Collect Measurements $\mathbf{Y}_{(1)} = \mathbf{\Phi M S}$
Solve OP: $\{\widehat{\mathbf{L}}, \widehat{\mathbf{C}}\} = \text{argmin}_{\boldsymbol{L},\boldsymbol{C}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{C}\|_{1,2}$ s.t. $\|\mathbf{Y}_{(1)} - \boldsymbol{L} - \boldsymbol{C}\|_F \le \varepsilon_1$
Estimate $\widehat{\mathbf{L}}_{(1)} = \widehat{\mathbf{U}} \mathcal{D}_\alpha(\widehat{\mathbf{\Sigma}}) \widehat{\mathbf{V}}^*$, where $\widehat{\mathbf{L}} = \widehat{\mathbf{U}} \widehat{\mathbf{\Sigma}} \widehat{\mathbf{V}}^*$ and $\mathcal{D}_\alpha(\widehat{\mathbf{\Sigma}})$ preserves entries of $\mathbf{\Sigma}$ larger than α.
Step 2. Let $\widehat{\mathcal{L}}_{(1)}$ be the linear subspace spanned by col's of $\widehat{\mathbf{L}}_{(1)}$
Set $\mathbf{P}_{\widehat{\mathcal{L}}_{(1)}^\perp} \triangleq \mathbf{I} - \mathbf{P}_{\widehat{\mathcal{L}}_{(1)}}$ and Collect $\mathbf{Y}_{(2)} = \mathbf{\Psi}\, \mathbf{P}_{\widehat{\mathcal{L}}_{(1)}^\perp}(\mathbf{\Phi M})$
Output: $\widehat{\mathcal{I}}_{\mathbf{C}} = \{i : \mathbb{1}(\|(\mathbf{Y}_{(2)})_{:,i}\|_2 > \varepsilon_2) = 1\}$

The authors graciously acknowledge the support from NSF Award CCF-1217751 and DARPA Young Faculty Award N66001-14-1-4047.

II. Performance Guarantees

We first introduce two terminologies: (i) Given the compact SVD $\mathbf{L} = \mathbf{U\Sigma V}^*$, $\mathbf{L}$ is said to satisfy the **column incoherence property** with parameter $\mu_{\mathbf{V}} \in [1, n_{\mathbf{L}}/r]$ if $\max_{j\in[n_2]} \|\mathbf{V}^* \mathbf{e}_j\|_2^2 \le \mu_{\mathbf{V}} \frac{r}{n_{\mathbf{L}}}$, where $\{\mathbf{e}_j\}$ are canonical basis vectors for $\mathbb{R}^{n_2}$; and (ii) A random matrix $\mathbf{\Phi} \in \mathbb{R}^{m\times n}$ is said to satisfy the **distributional JL property** if $\Pr\big(\,\big|\, \|\mathbf{\Phi v}\|_2^2 - \|\mathbf{v}\|_2^2 \,\big|\, \ge \varepsilon \|\mathbf{v}\|_2^2 \,\big) \le 2e^{-mf(\varepsilon)}$ for any fixed $\mathbf{v} \in \mathbb{R}^n$ and $\varepsilon \in (0,1)$, where $f(\varepsilon) > 0$ is a constant depending only on ε that is specific to the distribution of $\mathbf{\Phi}$.

Motivated from the noiseless case in [9], we state the *structural conditions* for noisy observations as follows: (**d1**) $\text{rank}(\mathbf{L}) = r < \min\{n_1, n_2\}$; (**d2**) $\mathbf{L}$ has $n_{\mathbf{L}} = n_2 - k$ nonzero columns; (**d3**) $\mathbf{L}$ satisfies the *column incoherence property* with parameter $\mu_{\mathbf{V}}$; (**d4**) the condition number of $\mathbf{L}$ satisfies $\kappa = \frac{\sigma_1(\mathbf{L})}{\sigma_r(\mathbf{L})} < \infty$; and (**d5**) $\mathbf{C}$ has $|\mathcal{I}_{\mathbf{C}}| = k$ nonzero columns, where $\mathcal{I}_{\mathbf{C}} = \{i \in [n_2] : \|\mathbf{P}_{\mathcal{L}^\perp} \mathbf{C}_{:,i}\|_2 > \tau_1 \|\mathbf{C}_{:,i}\|_2\}$ for some constant $\tau_1 \in (0,1)$.

Due to the existence of noise, we required further structural conditions of $\mathbf{N}$: (**n1**) $\sigma_r(\mathbf{L}) > \frac{90\sqrt{2\gamma}}{\tau_1} n_2 \eta_{\mathbf{N}}$; and (**n2**) $\min_{i\in\mathcal{I}_{\mathbf{C}}} \|\mathbf{C}_{:,i}\|_2 > \tau_2 \eta_{\mathbf{N}}$ for some constant τ_2, where $\eta_{\mathbf{N}} = \max_{j\in[n_2]} \|\mathbf{N}_{:,j}\|_2$. These conditions hold trivially for noiseless case when $\mathbf{N} = \mathbf{0}$. Then the main result is provide as follows.

Theorem II.1 (Accurate Recovery via RACOS-N)**.** *For model* (1)*, suppose that* $\mathbf{L}$ *and* $\mathbf{C}$ *satisfy (**d1**)-(**d5**) with* $k \le \frac{n_2}{3(1+1024\ r\mu_{\mathbf{V}})}$*. Let the measurement matrices* $\mathbf{\Phi}$ *and* $\mathbf{\Psi}$ *satisfy the distributional JL property, and for a fixed* $\delta \in (0,1)$*, suppose that the column subsampling parameter* γ*, and the row and column sampling parameters* m *and* q*, respectively, satisfy*

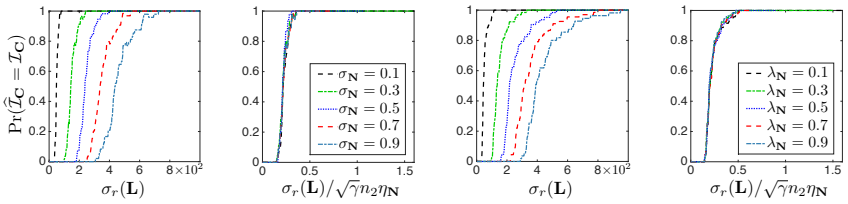
$$\gamma \ge \max\left\{ \frac{200\log(\frac{6}{\delta})}{n_{\mathbf{L}}}, \frac{600(1+1024r\mu_{\mathbf{V}})\log(\frac{6}{\delta})}{n_2}, \frac{10r\mu_{\mathbf{V}}\log(\frac{6r}{\delta})}{n_{\mathbf{L}}} \right\},$$

$$m \ge \frac{5(r+1)+\log(2n_2)+\log(\frac{2}{\delta})}{f(1/4)}, \text{ and } q \ge \frac{4\log(\frac{2n_2}{\delta})}{f(1/4)}.$$

Further suppose that $\mathbf{N}$ *satisfies (**n1**) and (**n2**), where* τ_2 *satisfies* $\tau_1\tau_2 > 6(\beta+1)(\tau_1/4+1) + 90\sqrt{6\gamma}\beta\kappa n_2$ *with* $\beta > \sqrt{3}$*, and* λ *in OP satisfies* $\lambda = \frac{3\sqrt{1+1024\mu_{\mathbf{V}} r}}{14\sqrt{\check{n}_2}}$*, where* $\check{n}_2$ *is the number of columns of* $\mathbf{S}$*. Then there exist constants* α *and* ε_2 *satisfying* $18\gamma n_2\eta_{\mathbf{N}} < \alpha < 54\gamma n_2\eta_{\mathbf{N}}$ *and* $\max_{j\in\mathcal{I}_{\mathbf{L}}} \|\mathbf{\Psi P}_{\widehat{\mathcal{L}}_{(1)}^\perp}(\mathbf{\Phi M}_{:,j})\|_2 < \varepsilon_2 < \min_{i\in\mathcal{I}_{\mathbf{C}}} \|\mathbf{\Psi P}_{\widehat{\mathcal{L}}_{(1)}^\perp}(\mathbf{\Phi M}_{:,i})\|_2$*, then with probability at least* $1-3\delta$*, we have simultaneously:*

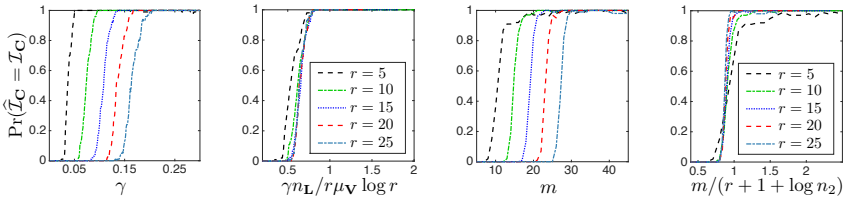
(**C1**) *RACOS-N correctly identifies outliers (i.e.,* $\widehat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}}$*), and*
(**C2**) *the total number of measurements collected is no greater than* $\big(\big(\frac{3}{2}\big)\gamma m + q\big) n_2$.

Theorem II.1 guarantees that RACOS-N succeeds with an effective sampling rate $\frac{\#\text{obs}}{n_1 n_2} = \mathcal{O}\Big(\frac{(r+\log n_2)(n_2/n_{\mathbf{L}})\mu_{\mathbf{V}} r\log r}{n_1 n_2} + \frac{\log n_2}{n_1}\Big)$ w.h.p. Note that we provide a result for deterministic noise $\mathbf{N}$. Improved result can be obtained for random $\mathbf{N}$. Numerical evaluations are provided in Figure 1, 2, and 3 to justify the tightness of the bounds for our parameters and the improvement on computational cost over the full-size data model. We refer [9], [11] for real data experiments on salient image feature detection, and [12] for a full version of this paper with detailed analysis. We further refer extensions of the model to tensor outliers [13] and dictionary based outliers [14].



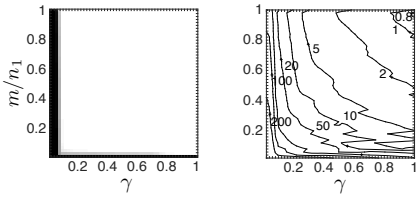
(a) (b) (c) (d)

Fig. 1. Demonstration of the probability of success $\Pr(\widehat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ versus the minimal singular value $\sigma_r(\mathbf{L})$ of $\mathbf{L}$ for zero-mean Gaussian noise under different choices of the variance $\sigma_{\mathbf{N}}$ (a and b) and zero-mean Laplace noise under different choices of the parameters $\lambda_{\mathbf{N}}$ (c and d). (b) and (d) provide the results with rescaling of $\sigma_r(\mathbf{L})$ by $\sqrt{\gamma} n_2 \eta_{\mathbf{N}}$. A trial is deemed a success if $\min_{i \in \mathcal{I}_{\mathbf{C}}} \| \mathbf{\Psi}\ \mathbf{P}_{\widehat{\mathcal{L}}_{(1)}^{\perp}} (\mathbf{\Phi} \mathbf{M}_{:,i}) \|_2 > \max_{i \in \mathcal{I}_{\mathbf{L}}} \| \mathbf{\Psi}\ \mathbf{P}_{\widehat{\mathcal{L}}_{(1)}^{\perp}} (\mathbf{\Phi} \mathbf{M}_{:,i}) \|_2$. We generate both the row sampling matrix $\mathbf{\Phi}$ and the row reduction matrix $\mathbf{\Psi}$ with i.i.d. $\mathcal{N}(0,1)$ entries. We fix $n_1 = 100$, $n_2 = 1000$, $q = 20$, $k = 0.2 n_2$, $n_{\mathbf{L}} = n_2 - k$, $\lambda = 0.4$, $r = 5$, $m = 0.3 n_1$ and $\gamma = 0.2$. We generate two random matrices $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_{\mathbf{L}} \times r}$ with i.i.d. $\mathcal{N}(0,1)$ entries, and take $\mathbf{L}_0 = [\mathbf{U}\mathbf{V}^T\ \mathbf{0}_{n_1 \times k}]$. Then let $\mathbf{L} = \frac{\sigma_r(\mathbf{L})}{\sigma_r(\mathbf{L}_0)} \mathbf{U}_0 \mathbf{\Sigma}_0 \mathbf{V}_0^T$, where $\mathbf{U}_0 \mathbf{\Sigma}_0 \mathbf{V}_0^T$ is SVD of $\mathbf{L}_0$, $\sigma_r(\mathbf{L}_0) = (\mathbf{\Sigma}_0)_{rr}$ is the minimal singular value of $\mathbf{L}_0$, and $\sigma_r(\mathbf{L})$ is a parameter to control the singular values of $\mathbf{L}$. In panel (a), we observe that as $\sigma_{\mathbf{N}}$ increases, the threshold of $\sigma_r(\mathbf{L})$ for correct identification of outlier columns with high probability also increases, as we expect. On the other hand, when we rescale $\sigma_r(\mathbf{L})$ by $\sqrt{\gamma} n_2 \eta_{\mathbf{N}}$ in panel (b), all curves corresponding to different values of $\sigma_{\mathbf{N}}$ are aligned together. In addition, when the ratio $\frac{\sigma_r(\mathbf{L})}{\sqrt{\gamma} n_2 \eta_{\mathbf{N}}}$ goes beyond 1, the probability of correct outlier detection is 1, which justifies our assumption (**n1**) in this case. Analogous results are observed for Laplace noise as well.



(a) (b) (c) (d)

Fig. 2. Demonstration of the probability of success $\Pr(\widehat{\mathcal{I}}_{\mathbf{C}} = \mathcal{I}_{\mathbf{C}})$ versus column subsample parameter γ (a and b) and row sampling parameter m (c and d) for noisy observations under different settings of rank r of $\mathbf{L}$. (b) and (d) provide the results with rescaling of γ by $\frac{r \mu_{\mathbf{V}} \log(r)}{n_{\mathbf{L}}}$ and m by $r + 1 + \log k$ respectively. We fix $\mathbf{N}$ as Gaussian noise with i.i.d. entries with $\sigma_{\mathbf{N}} = 0.01$. We generate $\mathbf{L} = [\mathbf{U}\mathbf{V}^T\ \mathbf{0}_{n_1 \times k}]$ and $\mathbf{C} = [\mathbf{0}_{n_1 \times n_{\mathbf{L}}}\ \mathbf{W}]$, where $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_{\mathbf{L}} \times r}$ have i.i.d. $\mathcal{N}(0,1)$ entries and $\mathbf{W} \in \mathbb{R}^{n_1 \times k}$ has i.i.d. $\mathcal{N}(0, r)$ entries. When r increases, the column subsampling parameter γ also needs to increase for correct outlier identification with high probability. If we normalize γ with $\frac{r \mu_{\mathbf{V}} \log r}{n_{\mathbf{L}}}$, which is generally the dominating term, then all curves corresponding to different ranks r align together, as shown in panel (b). Further, high probability of success is achieved when the ratio $\gamma / \frac{r \mu_{\mathbf{V}} \log r}{n_{\mathbf{L}}} > 1$. Analogously, increasing m facilitates the accurate recovery for increasing r, and the ratio $m/(r + 1 + \log n_2) > 1$ facilitates correct recovery with high probability, as shown in panel (d).



(a) (b)

Fig. 3. Demonstration of the performance using different combinations of m and γ for noisy observations via (a) phase transition and (b) contour plot of timing evaluation of OP. We fix $n_1 = 500$, $n_2 = 1000$, $k = 0.2 n_2$, $n_{\mathbf{L}} = n_2 - k$, $r = 10$, and $\lambda = 0.4$, and generate $\mathbf{L}$, $\mathbf{C}$, and the Gaussian noise $\mathbf{N}$ in the same way in Figure 2. The pair $(m, \gamma) = (500, 1)$ corresponds to operating on the full-size data matrix $\mathbf{M}$. We first provide the "phase transition" behavior for all combinations of m and γ, and a fixed $\lambda = 0.5$ in OP. Then we record the CPU execution time of Algorithm 1. The values on contour lines are the speed-ups of algorithm compared with the full size model, *i.e.* $(m, \gamma) = (500, 1)$. We can see that our approach shows significant advantage in terms of computational efficiency over the full data model when m and γ are small. For example, when $(m/n_1, \gamma) = (0.1, 0.1)$, our approach is > 100 times faster than that using the full data. Another interesting observation is that the full size model $(m, \gamma) = (500, 1)$ is not the slowest here, while the nearly full size model is the slowest. This is because in the full data model, we do not need to construct the random projection matrices and the corresponding projection operations. In real data applications, such as the salient image feature detection, speedup of over 100 times can be achieved with comparable performances [9].

REFERENCES

[1] B. Mehta and W. Nejdl, "Attack resistant collaborative filtering," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 75–82.

[2] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4. ACM, 2004, pp. 219–230.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. CVPR*, 2012, pp. 853–860.

[5] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[7] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3047–3064, 2012.

[8] M. Soltanolkotabi and E. Candes, "A geometric analysis of subspace clustering with outliers," *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

[9] X. Li and J. Haupt, "Identifying outliers in large matrices via randomized adaptive compressive sampling," *Trans. Signal Processing*, vol. 63, no. 7, pp. 1792–1807, 2015.

[10] ——, "A refined analysis for the sample complexity of adaptive compressive outlier sensing," in *IEEE Workshop on Statistical Signal Processing*, 2016.

[11] ——, "Locating salient group-structured image features via adaptive compressive sensing," in *GlobalSIP*, 2015.

[12] ——, "Robust low-complexity randomized methods for locating outliers in large matrices," *arXiv preprint arXiv:1612.02334*, 2016.

[13] J. Ren, X. Li, and J. Haupt, "Robust pca via tensor outlier pursuit," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 1744–1749.

[14] X. Li, J. Ren, Y. Xu, and J. Haupt, "An efficient dictionary based robust pca via sketching," Technical Report, Tech. Rep., 2016.