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 $M \in \mathbb{R}^{n_1 \times n_2}$, we assume that

$$M = L + C + N,$$

(1)

where $L$ is a rank-$r$ matrix, $C$ is a column-sparse matrix with $k$ nonzero columns that are interpreted as "outliers" of the subspace spanned by columns of $L$, and $N$ is an additive noise. Our goal is to identify the locations $I_C$ of the nonzero columns of $C$, 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], we state the structural conditions for noisy observations as follows: (d1) rank($L$) = $r < \min(n_1, n_2)$; (d2) $L$ has $n_L = n_2 - k$ nonzero columns; (d3) $L$ satisfies the column incoherence property with parameter $\mu_L$; (d4) the condition number of $L$ satisfies $\kappa = \frac{\|L\|_2}{\|L\|_F} < \infty$; and (d5) $C$ has $|I_C| = k$ nonzero columns, where $I_C = \{i \in [n_2] : \|P_{\ell_i} \cdot C(i, \cdot)\|_2 > \gamma_{n_1} \|C\|_2\}$ for some constant $\gamma_1 \in (0, 1)$.

Due to the existence of noise, we required further structural conditions of $N$: $n_1 \sigma_1(L) > \frac{\ka}{\sqrt{n_2}}$; and (n2) $\min_{i \in I_C} \|C(i, \cdot)\|_2 > \gamma_{n_1} \|C\|_2$ for some constant $n_2$, where $\eta_n = \max_{i \in [n_2]} \|n_i\|_2$. These conditions hold trivially for noiseless case when $N = 0$. Then the main result is provide as follows.

Theorem II.1 (Accurate Recovery via RACOS-N). For model (1), suppose that $L$ and $C$ satisfy (d1)–(d5) with $k \leq \frac{\ka}{\sqrt{n_2} \gamma_{n_1} \|n_i\|_2}$. Let the measurement matrices $\Phi$ and $\Psi$ satisfy the distributional JL property, and for a fixed $\delta \in (0, 1)$, suppose that the column subsampling parameter $\gamma_1$, and the row and column sampling parameters $m$ and $q$, respectively, satisfy

$$\gamma_1 \geq \max \left\{ \frac{200 \log \left( \frac{\ka}{\sqrt{n_2}} \right)}{n_2}, \frac{600(1+1024\nu \eta)}{n_2}, \frac{10 \nu \eta}{n_2}, \frac{\log \left( \frac{\ka}{\sqrt{n_2}} \right)}{n_2}, \frac{\log \left( \frac{\ka}{\sqrt{n_2}} \right)}{n_2} \right\},$$

and

$$m \geq \frac{5(n+1+\log(2n_2) + \log(2q))}{f(\eta)} + \frac{4 \log \left( \frac{\ka}{\sqrt{n_2}} \right)}{f(\eta)},$$

Further suppose that $N$ satisfies (n1) and (n2), where $\gamma_2$ satisfies $\gamma_1 \gamma_2 > \beta(1+\delta)(1+\eta_1) + 900 \sqrt{3} \eta_2 \tau_1 \xi_1$ and $\delta$ in OP satisfies $\lambda < \frac{1}{\sqrt{1+\eta_2} \tau_2}$.

Then we have that $\eta_{n_1} \leq \frac{\ka}{\sqrt{n_2}}$ and $\max_{i \in I_C} \|P_{\ell_i} \cdot C(i, \cdot)\|_2 < \varepsilon_2 < \min_{i \in I_C} \|P_{\ell_i} \cdot C(i, \cdot)\|_2$, with probability at least $1 - 3\delta$, we have simultaneously:

(C1) RACOS-N correctly identifies outliers (i.e., $\hat{I}_C = I_C$), and (C2) the total number of measurements collected is no greater than $(\frac{1}{2} \gamma_1 n + q) n_2$.

Theorem II.1 guarantees that RACOS-N succeeds with an effective sampling rate $\frac{\log \left( \frac{\ka}{\sqrt{n_2}} \right)}{\log \left( \frac{\ka}{\sqrt{n_2}} \right)} = \Theta \left( \frac{\log \left( \frac{\ka}{\sqrt{n_2}} \right)}{\log \left( \frac{\ka}{\sqrt{n_2}} \right)} \right)$ w.h.p. Note that we provide a result for deterministic noise. Improved result can be obtained for random $N$. 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].

The authors graciously acknowledge the support from NSF Award CCF-1217751 and DARPA Young Faculty Award N66001-14-1-4047.
Fig. 1. Demonstration of the probability of success $\Pr(\tilde{C}_L = \tilde{C}_G)$ versus the minimal singular value $\sigma_N$ of $L$ for zero-mean Gaussian noise under different choices of the variance $\sigma_N$ (a and b) and zero-mean Laplace noise under different choices of the parameters $\lambda_N$ (c and d). (b) and (d) provide the results with rescaling of $\sigma_N$ by $\sqrt{n_2 N}$. A trial is deemed a success if $\min_{i \in \{0,1\}} \| \tilde{U}_{ML} \Phi_{M,i} \|_2$ is greater than $\max_{i \in \{0,1\}} \| \tilde{U}_{ML} \Phi_{M,i} \|_2$. We generate both the row sampling matrix $\Phi$ and the row reduction matrix $\Psi$ with i.i.d. $\mathcal{N}(0,1)$ entries. We fix $n_1 = 100, n_2 = 1000, q = 20, k = 0.2 n_2, n_L = n_2 - k, \lambda = 0.4, r = 5, m = 0.3 n_1$, and $\gamma = 0.2$. We generate two random matrices $U \in \mathbb{R}^{n_1 \times k}$ and $V \in \mathbb{R}^{n_1 \times r}$ with i.i.d. $\mathcal{N}(0,1)$ entries, and take $L_0 = [U^T \! 0_{n_1 \times k}]$. Then let $L = \sigma_N \tilde{L}_0 U_0 \Sigma_0 V_0^T$, where $U_0 \Sigma_0 V_0^T$ is SVD of $L_0$. We then record the CPU execution time of Algorithm 1. The phase transition behavior for all combinations of $M, G$ and $\sigma_N$ is not captured. Analogously, increasing $\sigma_N$ takes $L_0$ to be the “phase transition” behavior for all combinations of $M, G$ and $\sigma_N$. Then let $L = \sigma_N \tilde{L}_0 U_0 \Sigma_0 V_0^T$, where $U_0 \Sigma_0 V_0^T$ is SVD of $L_0$. Then we record the CPU execution time of Algorithm 1. The phase transition behavior for all combinations of $M, G$ and $\sigma_N$ is not captured.

Fig. 2. Demonstration of the probability of success $\Pr(\tilde{C}_C = \tilde{C}_G)$ versus column subsample parameter $\gamma$ (a and b) and row sampling parameter $m$ (c and d) for noisy observations under different settings of rank $\mathcal{R}$ of $L$. (b) and (d) provide the results with rescaling of $\gamma$ by $\frac{\gamma}{n_2}$ and $m$ by $r + 1 + \log k$ respectively. We fix $N$ as Gaussian noise with i.i.d. entries with $\sigma_N = 0.01$. We generate $L = [U^T \! 0_{n_1 \times k}]$ and $C = [0_{m \times n_1 \times k}]$, where $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_1 \times r}$ have i.i.d. $\mathcal{N}(0,1)$ entries and $W \in \mathbb{R}^{n_1 \times k}$ has i.i.d. $\mathcal{N}(0,1)$ entries. We can see that our approach shows significant advantage in terms of computational efficiency over the full data model when $m$ and $\gamma$ 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].

Fig. 3. Demonstration of the performance using different combinations of $m$ and $\gamma$ for noisy observations via (a) phase transition and (b) contour plot of timing evaluation of OP. We fix $n_1 = 500, n_2 = 1000, k = 0.2 n_2, n_L = n_2 - k, r = 10, \lambda = 0.4$, and generate $C_L, C$, 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 $\gamma$, and a fixed $\lambda = 0.5$ in OP. Then we record the CPU execution time of Algorithm 1. The phase transition behavior for all combinations of $m$ and $\gamma$ is 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


