Robust Low-Complexity Methods for Matrix Column Outlier Identification

Xingguo Li and Jarvis Haupt
Department of Electrical and Computer Engineering
University of Minnesota Twin Cities, Minneapolis, MN, 55455
email: {lixx1661, jdhaupt}@umn.edu

Abstract—This paper examines the problem of locating outlier columns in a large, otherwise low-rank matrix, in settings where the data are noisy, or where the overall matrix has missing elements. We propose an efficient 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 for each task. Comprehensive numerical experimental results are provided to validate the theoretical bounds and demonstrate the computational efficiency of the proposed algorithm.

I. INTRODUCTION

This paper examines two variants of a matrix outlier identification problem. Let $M \in \mathbb{R}^{n_1 \times n_2}$ represent a data matrix, which we assume is approximately low-rank, but corrupted by (nominally few) outlier columns. Our ultimate goal is to identify the locations of the nonzero outlier columns, without necessarily identifying the “inlier” vectors whose columns comprise the low-rank part of $M$ (or the subspace they span), and our particular interest is in doing so in settings where our observations of $M$ may either be contaminated by additive noise, or where only a subset of elements of $M$ are available. In both cases, our aim is to develop computationally-efficient methods that are readily deployable in scenarios where the dimensions $n_1$ and $n_2$ are very large relative to the rank of the “inlier” matrix and the number of outliers. This work builds upon our own recent work [1], [2] (and an application-based extension [3]), which analyzed a two-step randomized projection and downsampling algorithm (applicable in noise free settings, and where the data do not suffer from missing elements), establishing provable recovery guarantees and the significant reductions in sampling and computational complexities of our approach relative to “full data” variants.

Our overall investigation is motivated by a wide class of “big data” applications where the outliers themselves are of interest, such as when identifying malicious responses in collaborative filtering applications [4], finding anomalous patterns in network traffic [5], and many others [6]–[8]. These and related problems fall within the broad scope of so-called robust principal component analysis (PCA) tasks, for which a body of recent work leverages low-dimensional models using convex inference methods. Among these, [9], [10] examine robust PCA problems based on entry-wise sparse corruptions, while [11]–[16] propose and analyze methods applicable when outliers are present as entire columns. Despite the provable analytical successes of these methods, they can be computationally demanding when applied to very large data matrices and, more notably for our purposes here, these existing techniques are designed to identify or approximate the low-rank component of $M$ (or the subspace spanned by these “inlier” columns). Here, as alluded above, our interest is only in locating the outlier columns, and we seek inference procedures having both low sample and implementation complexities (e.g., to obviate the need to store and process the full data matrix).

For completeness, we also note several recent works that leveraged randomized algorithms to improve the computational efficiency of robust subspace estimation procedures. A divide-and-conquer type algorithm was proposed using random division of data and parallel computation [17]. A randomized downsampling-selection based algorithm was proposed in [18]. Random projection based algorithms were also popular [19], [20] due to their efficiencies and robustness to random noise. Subsequent works leveraged similar dimensionality reduction methods as in our works [1], [2] to devise efficient subspace recovery procedures [21], [22].

A. Overview of Our Contribution

As alluded above, our major contributions in this paper are extensions of the approach of [1], [2] to settings where the data is corrupted by additive noise, or where the available data are incomplete. In the noisy setting, we analyze a randomized sampling and inference procedure that successfully locates outliers w.h.p. using an effective sampling rate of $\frac{\#obs}{n_1 n_2} = \mathcal{O} \left( \frac{(r+\log n_2)(n_2/n_L)\mu_V \mu_L \log r + \log n_2}{n_1 n_2} \right)$; in missing-data settings, we present a procedure that succeeds w.h.p. using an effective sampling rate $\frac{\#obs}{p n_1 n_2} = \mathcal{O} \left( \frac{p n_1 \log^2 n_2}{p n_1 n_2} \right)$, where $n_L$ is the number of inliers, $p$ is observation rate in the missing-data setting, and $\mu_V$ and $\mu_L$ are incoherence parameters.

B. Algorithm for Noisy Observations

We first suppose that

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

where $L$ is a rank-$r$ matrix, $C$ is a column-sparse matrix with $k$ nonzero columns, and $N$ is a matrix of additive noise. The key insight in our two-step approach can be described as follows. In the first step, we apply an existing robust PCA approach designed to be robust to column outliers – called Outlier...
Pursuit (OP) [11] – to a matrix comprised of a small random subset of columns of $\Phi M$. This results, in part, in an estimate of the low-rank component $\Phi L$ of $\Phi M$, and we identify the subspace spanned from this estimate by the singular vectors of $\Phi L$ corresponding to singular values above a specified threshold (this serves to mitigate the effects of the noise in the subspace estimate). Specifically, let the (compact) SVD of $\Phi$ be $\Phi = \hat{U}\Sigma\hat{V}^\ast$, where $\hat{U} \in \mathbb{R}^{n_1 \times r}$ and $\hat{V} \in \mathbb{R}^{n_2 \times r}$ have orthonormal columns, $\Sigma = \text{diag}(\{\hat{\sigma}_i\})_{1 \leq i \leq \min(m, n_2)}$, and $n_2$ is the number of columns of $S$. By choosing a constant $\alpha$, we then apply a singular value thresholding operation defined as $D_\alpha (\Sigma) = \text{diag}(\{f(\hat{\sigma}_i, \alpha)\})_{1 \leq i \leq \min(m, n_2)}$, where $f(\hat{\sigma}_i, \alpha) = \hat{\sigma}_i$ if $\hat{\sigma}_i > \alpha$; or 0 otherwise. Then the estimate of the low-rank matrix is $\hat{L}(1) = \hat{U}D_\alpha (\Sigma)\hat{V}^\ast$.

Then, a second step incorporates into the sampling operation a composition of an orthogonal projection onto the orthogonal complement of the low-rank component $\hat{L}(1)$, indexed also by the rows in $I_j$. A column is recognized as an outlier if its energy after this orthogonal projection is nonzero. We call the algorithm RACOS for incomplete observation (RACOS-I), and summarize it in Algorithm 2.

**Algorithm 1** Robust Adaptive Compressive Outlier Sensing for Noisy Observations (RACOS-N)

**Input:** $M$, $\gamma \in (0, 1)$, $\lambda$, $\alpha$, $\epsilon_1, \epsilon_2 > 0$, and $q, m \in [n_1]$

**Initialize:** $\Phi \in \mathbb{R}^{m \times n_1}$, $\Psi \in \mathbb{R}^{q \times m}$ and $S = I_{S, S}$, where $S = \{j \in [n_2] : S_j \overset{\text{id}}{\sim} \text{Bernoulli}(\gamma) = 1\}$

**Step 1**

Collect Measurements: $Y(1) = \Phi MS$

Solve OP: $\{\hat{L}, \hat{C}\} = \text{argmin}_{L,C} \|L\|_* + \lambda\|C\|_{1,2}$

Estimate: $\hat{L}(1)$ by singular value thresholding operation on $L$, i.e. $\hat{L}(1) = \hat{U}D_\alpha (\Sigma)\hat{V}^\ast$

**Step 2**

Let: $\tilde{L}(1)$ be the linear subspace spanned by col’s of $\hat{L}(1)$

Set: $P_{\tilde{L}(1)} \triangleq I - P_{\tilde{L}(1)}$

Collect Measurements: $Y(2) = \Psi P_{\tilde{L}(1)}(\Phi M)$

Set: $\tilde{z}_i = \|Y(2)_{:,i}\|_2$, if $\|Y(2)_{:,i}\|_2 > \epsilon_2$

Output: $\tilde{I}_C = \{i : \tilde{z}_i \neq 0\}$

**C. Algorithm for Incomplete Observations**

We also consider variants of the outlier identification problem when the matrix $M$ has missing elements, i.e.,

$$M = P_{\Omega}(L + C),$$

where $P_{\Omega}$ masks its arguments that are not in the index set $\Omega \subseteq [n_1] \times [n_2]$ (where $[n] \overset{\triangleq}{=} \{1, 2, \ldots, n\}$ for an integer $n$). Here $\Omega$ follows an independent Bernoulli model with parameter $p$, i.e., each entry of $[n_1] \times [n_2]$ is contained in $\Omega$ independently with probability $p$.

Here we employ an analogous approach as in the noisy case. Namely, we operate throughout on a column-wise compressed matrix $\Phi M$, but consider specifically the case where $\Phi$ is a row submatrix of the $n_1 \times n_1$ identity matrix. In the first step, we apply an existing robust PCA approach – called Manipulator Pursuit (MP) [15], [23] – to a small random subset of columns of $\Phi M$. This results in an estimate of the low-rank matrix $\Phi L$, denoted by $\hat{L}(1)$. A trimming procedure is applied before MP by throwing away some entries randomly, which provides improved sampling complexity [15].

Then, the second step entails a missing data variant of the orthogonal projection discussed above. Let $I_j \subset [n_1]$ denote the indices of the observed elements of the $j$-th column of $\Phi M$. Then, for each $j \in [n_2]$, we project the observed subvector of the $j$-th column of $\Phi M$ onto the orthogonal complement of the column space of a row submatrix of $\hat{L}(1)$, indexed also by the rows in $I_j$. A column is recognized as an outlier if its energy after this orthogonal projection is nonzero. We call the algorithm RACOS for incomplete observation (RACOS-I), and summarize it in Algorithm 2.

**Algorithm 2** Robust Adaptive Compressive Outlier Sensing for Incomplete Observations (RACOS-I)

**Input:** $M$, $\Omega$, $\gamma_1, \gamma_2 \in (0, 1)$, $\rho$ and $\lambda > 0$

**Initialize:** $\Phi = I_{S_1}, S = I_{S, S}$, where $S_1 = \{i \in [n_1] : S_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\gamma_1) = 1\}$, $m = |S_1|$, $S_2 = \{j \in [n_2] : S_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\gamma_2) = 1\}$, $n_2 = |S_2|$

**Step 1**

Collect Measurements: $Y(1) = \Phi MS$

**for** $j = 1$ to $n_2$ (Trimming)

if number of observed entries of $(Y(1))_{:,j} > \rho m$

Select: $pm$ entries of $(Y(1))_{:,j}$ uniformly randomly

Set: The rest entries of $(Y(1))_{:,j}$ unobserved

**end for**

Set: $\Omega(1)$ be the set of observed entries of $Y(1)$

Solve MP: $\{\hat{L}(1), \hat{C}(1)\} = \text{argmin}_{L,C} \|L\|_* + \lambda\|C\|_{1,2}$

Estimate: $\hat{I}(1) = P_{\Omega(1)}(L + C)$

**Step 2**

**for** $j = 1$ to $n_2$

Let: $\tilde{L}_{I_j}$ be subspace spanned by col’s of $(\hat{L}(1))_{I_j,:}$

Set: $P_{\tilde{L}_{I_j}} \triangleq I - P_{\tilde{L}_{I_j}}$ and $\tilde{z}_j = \|P_{\tilde{L}_{I_j}}(\Phi M)_{I_j,:}\|_2$

**end for**

Output: $\tilde{I}_C = \{i : \tilde{z}_i \neq 0\}$

**D. A Note on Notation**

We denote $P_{\Omega}(X) = UU^TX$ and $P_{\Omega}^\perp(X) = (I - UU^T)X$ as projection operations that project a matrix $X$ onto the column space and the orthogonal complement of column space of $L$ respectively. MATLAB-inspired notation is used to denote submatrices; e.g., $I_{S,:}$ (or $I_{S,\ast}$) is used to denote the submatrix formed by extracting rows (or columns) of $I$ indexed by $S$. Likewise, we use $X_{:,j}$ to denote the $j$-th column of $X$.

**II. MAIN RESULTS**

We start with introducing two important properties on which our recovery guarantees are based. It is well-known that the decomposition of a matrix into a low-rank component and a sparse component is a ill-posed problem in general (for
example, a matrix with only one non-zero entry is both a low-rank and sparse matrix). The first property outlined below is a widely adopted notion of “incoherence” in the literature of robust PCA [9–11], used to overcome identifiability issues.

**Definition II.1 (Row and Column Incoherence Properties).** Let $L \in \mathbb{R}^{n_1 \times n_2}$ be a rank $r$ matrix with at most $n_L \leq n_2$ nonzero columns. Given the compact SVD $L = U \Sigma V^*$, $L$ is said to satisfy the **row incoherence property** with parameter $\mu U \in [1, n_1/r]$ if $\max_{i \in [n_1]} ||U^*e_i||_2 \leq \mu U \eta n_2$, where $\{e_i\}$ are canonical basis vectors for $\mathbb{R}^{n_1}$. Likewise, $L$ is said to satisfy the **column incoherence property** with parameter $\mu V \in [1, n_2/r]$ if $\max_{j \in [n_2]} ||V^*e_j||_2 \leq \mu V \eta n_2$, where $\{e_j\}$ are canonical basis vectors for $\mathbb{R}^{n_2}$.

The second property is a criterion on the measurement matrix as following.

**Definition II.2 (Distributional Johnson-Lindenstrauss (JL) Property) [24].** A matrix $\Phi \in \mathbb{R}^{m \times n}$ is said to satisfy the **JL property** if for any fixed $v \in \mathbb{R}^{n}$ and any $\varepsilon \in (0, 1)$,

$$\Pr\left( \left( ||\Phi \eta v||_2^2 - ||v||_2^2 \right) \geq \varepsilon ||v||_2^2 \right) \leq 2e^{-mf(\varepsilon)}, \quad (3)$$

where $f(\varepsilon) > 0$ is a constant depending only on $\varepsilon$ that is specific to the distribution of $\Phi$.

**A. Guarantees for Noisy Observations**

Motivated by our work in the noise-free case [1], [2], we state the **structural conditions** for noisy observations as following (adapted from [1]):

**(d1)** rank$(L) = r$, **(d2)** $L$ has $n_L \leq n_2 - k$ nonzero columns, **(d3)** $L$ satisfies the column incoherence property with $\mu V$, **(d4)** the condition number of $L$ satisfies $\kappa = \frac{\sigma_1(L)}{\sigma_r(L)} < \infty$, and **(d5)** $C$ has $|I_C| = k$ nonzero columns, where $I_C \triangleq \{ i \in [n_2] : \|P_{C^c} \mathbf{C}_{i,:}\|_2 > \tau_1 \|\mathbf{C}_{i,:}\|_2 \}$ for some constant $\tau_1 \in (0, 1)$.

Conditions (d1)-(d3) are natural for $L$, and are similar to those imposed in our prior work [1], [2]. Condition (d4) assumes the well-conditioning of $L$. That the residuals of the outlier columns need to be sufficiently large in (d5) is somewhat intuitive due to the inexact estimate of noisy low-rank subspace. In our analysis, $\tau_1$ is an upper bound of the estimation error (in spectral norm) of the low-rank subspace; $\tau_1 = 0$ when $N = 0_{n_1 \times n_2}$. Thus (d5) is a natural extension of the analogous condition for the noise-free case [1], [2].

We also impose conditions on the noise to facilitate exact outlier detection. Indeed if, for example, $N$ has very large Euclidean norm in some inlier column, then we may confuse it with an outlier column. To avoid such unidentifiability, we impose conditions on the noise in terms of $L$ and $C$. For notational simplicity, we define $\eta_N = \max_{i \in [n_2]} \|N_{i,:}\|_2$. Then the structural conditions on $N$ are as following:

**(n1)** $\sigma_r(L) > \frac{100\sqrt{n_2} \eta n_2}{\eta_N}$, and **(n2)** $\min_{i \in I_C} \|C_{i,:}\|_2 > 2\eta_N \gamma$ for some constant $\gamma$.

where $\gamma \in (0, 1)$ is the column subsampling parameter (an input to Algorithm 1). Condition (n1) is akin to an SNR assumption, and ensures that all singular values of $L$ dominate the Euclidean norm of columns of $N$. Note that (n1) may seem somewhat restrictive, but our numerical evaluation indicates that $\sigma_r(L) = \Omega(\sqrt{n_2} \eta n_2)$ may be necessary for our approach. Condition (n2) is a structural one that any outlier column dominates the per-column noise. Note that conditions (n1) and (n2) hold somewhat trivially when $N = 0_{n_1 \times n_2}$.

In the next theorem, we state our main results for outlier identification for observation under model (1).

**Theorem II.1 (Accurate Recovery via RACOS-N).** Assume (1) where $L$ and $C$ satisfy the structural conditions (d1)-(d5) with the number of outliers $k$ satisfying $k < \frac{n_2}{3(1+1024\mu V)}$. Let the measurement matrices $\Phi$ and $\Psi$ be drawn from any distribution following (3), and for a fixed $\delta \in (0, 1)$, suppose that the column subsampling parameter $\gamma$, and the row and column sampling parameters $m$ and $n$, respectively, satisfy

$$\gamma \geq \max\left\{ \frac{200 \log\left(\frac{1}{\delta}\right)}{n_L}, \frac{600(1+1024\mu V) \log\left(\frac{1}{\delta}\right)}{n_L}, \frac{10\mu V \log\left(\frac{1}{\delta}\right)}{n_L}\right\},$$

$$m \geq \frac{5n(1)+\log(2n)+\log\left(\frac{1}{\delta}\right)}{(1/4)}, \quad q \geq 4\log(2n)/{(1/4)}.$$ Further suppose that $N$ satisfies the structural conditions (n1) and (n2), where the constant $\tau_2$ satisfies $\tau_1 \tau_2 > 6(\beta + 1)(\tau_1 / 4 + 1) + 90\sqrt{\beta} \kappa n_2$ with a constant $\beta > \sqrt{3}$, and the regularization parameter $\lambda$ in OP satisfies $\lambda = \frac{\mu U}{\sqrt{T} \cdot 10^{2\mu V}}$, where $n_2$ is the number of columns of $S$. Then there exist a singular value hard thresholding constant $\alpha$ satisfying $18\gamma n_2 \kappa \eta n_2 < \alpha < 54\gamma n_2 \kappa \eta n_2$ and an $\varepsilon_2$ satisfying

$$\max_{j \in [n]} \|\mathbf{P} E_{\mathbf{C},j} (\Phi \mathbf{M}_{j,:})\|_2 < \varepsilon_2 < \min_{i \in I_C} \|\Psi \mathbf{P} E_{\mathbf{C},i} (\Phi \mathbf{M}_{i,:})\|_2,$$

such that the following claims hold with probability at least $1 - 3\delta$: RACOS-N correctly identifies the salient columns of $C$ (i.e., $\hat{C} = C$), and the total number of measurements collected is no greater than $\left(\frac{3}{2}\right) \gamma m + q n_2$.

Note that the sufficient condition on $k$ is of the same order to the same order with OP [11] and noise-free ACOS [1], which can be as large as a fixed proportion of $n_2$ when both the rank $r$ and column coherence parameter $\mu V$ are small. Our results establishes that our approach succeeds w.h.p. with effective sampling rate

$$\frac{\text{#obs}}{n_1 n_2} = O\left( \frac{r + \log (n_2)}{n_1 n_2 \mu U \log(2r)} \right).$$

This may be much smaller than 1 when, e.g., $r \ll \min\{n_1, n_2\}$.

**B. Guarantees for Incomplete Observations**

For convenience, we denote $C_\Omega = P_\Omega (C)$. The **structural conditions** for the incomplete observation (2) are as following (adapted from [1]):

**(g1)** rank$(L) = r$, **(g2)** $L$ has $n_L$ nonzero columns, **(g3)** $L$ satisfies the row and column incoherence properties with parameters $\mu L$ and $\mu V$ respectively, and **(g4)** $C$ has $|I_C| = k$ nonzero columns, where $I_C \triangleq \{ j \in [n_2] : \forall \mathbf{I}^* \subset [n_1] \text{ with } |\mathbf{I}^*| \geq \frac{\mu L \log(2r)}{p}, \|\mathbf{P}_{\mathbf{I}^*} (C_\Omega)\|_2 > 0 \}.$

Note that the low-rank matrix $L$ needs to satisfy both column and row incoherence properties due to simultaneous column and row sampling procedure. Assumption (g4) is from the
fact that we only need to consider those observed entries in outlier columns. For missing entries in outlier columns, we will never be able to recover them exactly. In the next theorem, we provide the main result for of RACOS-I with trimming.

**Theorem II.2** (Accurate Recovery via RACOS-I with Trimming). Assume (2) where the components $L$ and $C$ satisfy the structural conditions (g1)-(g4). Assume $p$ and $k$ satisfy

$$p \geq P_r = C_p \left(1 + \frac{1}{\mu_L \log(2n_{1})/n_{1}}\right),$$

$$k \leq k_u = C_k \frac{1}{\mu_L \log(2n_{2})/n_{2}},$$

for some positive constants $C_p$ and $C_k$, where $\mu_L = \max(\mu_U, \mu_V).$ Given $\delta \in (0,1/2)$, if the row and column sampling parameters $\gamma_1$ and $\gamma_2$ satisfy

$$\gamma_1 \geq \max\left\{\frac{2r\mu_L \log(2r)}{n_{1} p}, \frac{8 \mu_L \log(2r)}{n_{1} p}, \frac{10 \mu_U \log(2r)}{n_{1}}, \frac{162p}{p}, \right\},$$

$$\gamma_2 \geq \max\left\{\frac{200 \mu_L \log(2r)}{n_{2}}, \frac{10 \mu_U \log(2r)}{n_{2}}, \frac{C_{\gamma_2}(2r)^{3/4}}{n_{2}}, \frac{200 \log(2)}{k_u}, \right\},$$

and the regularization parameter satisfies $\lambda = \frac{1}{2\delta} \sqrt{1 + \frac{r\mu_L k \log(n_{1} + n_{2})}{n_{1} p}}$, then the following hold simultaneously with probability at least $1 - 2\delta$: RACOS-I correctly identifies the salient columns of $C$, and the number of measurements collected is no greater than $\frac{3}{2}r_{1} n_{1} n_{2}$.

Consider $\varphi = O(1)$ for ease of discussion. Theorem II.2 for RACOS-I with trimming reduces the dimension of the matrix operated in MP to $O(\frac{1}{\mu_L} \log(2r)/n_{1} p)$ with an effective sampling rate $\#_{\text{obs}} = O\left(\frac{r\mu_L \log(2r)}{n_{1} p}\right)$. It is shown in [15] that (4) and (5) are close to information-theoretic (minimax) optimality, where the trimming step is crucial in the analysis. We omit all analysis, and refer to [25] for further details.

**III. EXPERIMENTAL EVALUATION**

In this section, we demonstrate explicitly via the numerical evaluation that the sample complexities we derived in the main results appear tight in practice. For all experiments below, refer to [25] for further details of data generation procedures.

**Noisy Observations.** A trial is deemed a success if we have

$$\min_{\ell \in \hat{I}_{C}} \|\Psi_{\ell, i} \Phi_{M, i}\|_2 > \max_{\ell \in \mathcal{I}_L} \|\Psi_{\ell, i} \Phi_{M, i}\|_2,$$

which implies that there exists a threshold $\varepsilon_2$ such that we can obtain accurate outlier identities. For the singular value thresholding, we choose $\alpha$ that preserves 99% of the sum of singular values, which performs well in our evaluations.

We first demonstrate that $\sigma_{C}(L) = \Omega(\sqrt{n_{1} n_{2}})$ in (n1) is necessary in practice using Gaussian noise matrix with i.i.d. $N(0, \frac{1}{2})$ entries. The empirical values of $Pr(\hat{I}_{C} = I_{C})$ (over 100 trials) for different choices of $\sigma_N$ are provided in Figure 1 (a,b). We observe that as $\sigma_N$ increases, the threshold of $\sigma_{C}(L)$ for correct identification of outlier columns w.h.p. also increases, as we expect. On the other hand, when we rescale $\sigma_{C}(L)$ by $\sqrt{n_{1} n_{2}}$, in panel (b), all curves corresponding to different values of $\sigma_N$ are aligned together. In addition, when the ratio $\frac{\sigma_{C}(L)}{\sigma_{N}(L)}$ goes beyond 1, the probability of correct outlier detection is 1, which justifies assumption (n1) here.

Next, we evaluate the bound of the column subsampling parameter $\gamma_1$ w.r.t. the rank $r$. We fix $N$ with i.i.d. $N(0,0.01)$ entries in the following discussion. The plots of $Pr(\hat{I}_{C} = I_{C})$ versus $\gamma_1$ for different values of $r$ are provided in Figure 1 (c,d). When $r$ increases, $\gamma_1$ also needs to increase for correct outlier identification w.h.p.. If we normalize $\gamma_1$ with $\frac{r \mu_L \log r}{n_{1} p}$ (the likely dominating term for $\gamma_1$), then all curves corresponding to different values of $r$ tend to align , as shown in panel (d). High probability of success is achieved when $\gamma_1 \geq \frac{r \mu_L \log r}{n_{1} p} > 1$. Analogous results holds for the row sampling parameter $m$.

**Incomplete Observations.** Analogous evaluations are performed for $\gamma_2$. The plots of $Pr(\hat{I}_{C} = I_{C})$ versus $\gamma_2$ are provided in Figure 2 (a,b). We observe that when $r$ increases, $\gamma_2$ also needs to increase for correct identification w.h.p. due to the positive dependence of $\gamma_1$ and $r$. When we rescale $\gamma_2$ by $\frac{r \mu_U \log r}{n_{2}}$ (the dominating term for $\gamma_2$), most curves corresponding to different $r$ align, facilitating high probability of recovery with the ratio exceeds 1. Similar results holds for the observation probability $p$ and column sampling rate $\gamma_2$.

**Timing Performance.** We finally examine the timing performance. The phase transition and the contour plot of timing evaluation are provided in Figure 2 (c,d). Note that white regions correspond to all successes and vice versa of the phase transition, and the values on contour lines are the speedups of algorithm compared with the full size model, i.e. $(m, \gamma) = (n_{1}, 1)$. We see that our approach shows significant computational advantage over the full data model when $m$ and $\gamma$ are small (e.g. $> 100$ times speedups when $(m/ n_{1}, \gamma) = (0.1, 0.1)$). Note that the full size model $(m, \gamma) = (n_{1}, 1)$ is not the slowest here, since there is no need to construct the random projection matrices and the corresponding projections.

**ACKNOWLEDGMENT**

The authors acknowledge support from NSF Award CCF-1217751 and the DARPA YFA, Award N66001-14-1-4047.


