Matrix Completion with Noisy Side Information

Kai-Yang Chiang[*] **Cho-Jui Hsieh** [†] **Inderjit S. Dhillon** [*]
[*] University of Texas at Austin [†] University of California at Davis
[*] {kychiang,inderjit}@cs.utexas.edu
[†] chohsieh@ucdavis.edu

Abstract

We study the matrix completion problem with side information. Side information has been considered in several matrix completion applications, and has been empirically shown to be useful in many cases. Recently, researchers studied the effect of side information for matrix completion from a theoretical viewpoint, showing that sample complexity can be significantly reduced given completely clean features. However, since in reality most given features are noisy or only weakly informative, the development of a model to handle a *general* feature set, and investigation of how much noisy features can help matrix recovery, remains an important issue. In this paper, we propose a novel model that balances between features and observations simultaneously in order to leverage feature information yet be robust to feature noise. Moreover, we study the effect of general features in theory and show that by using our model, the sample complexity can be lower than matrix completion as long as features are sufficiently informative. This result provides a theoretical insight into the usefulness of general side information. Finally, we consider synthetic data and two applications — relationship prediction and semi-supervised clustering — and show that our model outperforms other methods for matrix completion that use features both in theory and practice.

1 Introduction

Low rank matrix completion is an important topic in machine learning and has been successfully applied to many practical applications [22, 12, 11]. One promising direction in this area is to exploit the *side information*, or *features*, to help matrix completion tasks. For example, in the famous Netflix problem, besides rating history, profile of users and/or genre of movies might also be given, and one could possibly leverage such side information for better prediction. Observing the fact that such additional features are usually available in real applications, how to better incorporate features into matrix completion becomes an important problem with both theoretical and practical aspects.

Several approaches have been proposed for matrix completion with side information, and most of them empirically show that features are useful for certain applications [1, 28, 9, 29, 33]. However, there is surprisingly little analysis on the effect of features for general matrix completion. More recently, Jain and Dhillon [18] and Xu et al. [35] provided non-trivial guarantees on matrix completion with side information. They showed that if "perfect" features are given, under certain conditions, one can substantially reduce the sample complexity by solving a feature-embedded objective. This result suggests that completely informative features are extremely powerful for matrix completion, and the algorithm has been successfully applied in many applications [29, 37]. However, this model is still quite restrictive since if features are not perfect, it fails to guarantee recoverability and could even suffer poor performance in practice. A more general model with recovery analysis to handle noisy features is thus desired.

In this paper, we study the matrix completion problem with *general* side information. We propose a dirty statistical model which balances between feature and observation information simultaneously to complete a matrix. As a result, our model can leverage feature information, yet is robust to noisy features. Furthermore, we provide a theoretical foundation to show the effectiveness of our model. We formally quantify the quality of features and show that the sample complexity of our model depends on feature quality. Two noticeable results could thus be inferred: first, unlike [18, 35], given any feature set, our model is guaranteed to achieve recovery with at most $O(n^{3/2})$ samples in *distribution-free* manner, where n is the dimensionality of the matrix. Second, if features are reasonably good, we can improve the sample complexity to $o(n^{3/2})$. We emphasize that since $\Omega(n^{3/2})$ is the lower bound of sample complexity for distribution-free, trace-norm regularized matrix completion [32], our result suggests that even noisy features could asymptotically reduce the number of observations needed in matrix completion. In addition, we empirically show that our model outperforms other completion methods on synthetic data as well as in two applications: relationship prediction and semi-supervised clustering. Our contribution can be summarized as follows:

- We propose a dirty statistical model for matrix completion with *general* side information where the matrix is learned by balancing features and pure observations simultaneously.
- We quantify the effectiveness of features in matrix completion problem.
- We show that our model is guaranteed to recover the matrix with any feature set, and moreover, the sample complexity can be lower than standard matrix completion given informative features.

The paper is organized as follows. Section 2 states some related research. In Section 3, we introduce our proposed model for matrix completion with general side information. We theoretically analyze the effectiveness of features in our model in Section 4, and show experimental results in Section 5.

2 Related Work

Matrix completion has been widely applied to many machine learning tasks, such as recommender systems [22], social network analysis [12] and clustering [11]. Several theoretical foundations have also been established. One remarkable milestone is the strong guarantee provided by Candès et al. [7, 5], who proves that $O(n\text{polylog}n)$ observations are sufficient for exact recovery provided entries are uniformly sampled at random. Several work also studies recovery under non-uniform distributional assumptions [30, 10], distribution-free setting [32], and noisy observations [21, 4].

Several works also consider side information in matrix completion [1, 28, 9, 29, 33]. Although most of them found that features are helpful for certain applications [28, 33] and cold-start setting [29] from their experimental supports, their proposed methods focus on the non-convex matrix factorization formulation without any theoretical guarantees. Compared to them, our model mainly focuses on a convex trace-norm regularized objective and on theoretical insight on the effect of features. On the other hand, Jain and Dhillon [18] (also see [38]) studied an inductive matrix completion objective to incorporate side information, and followup work [35] also considers a similar formulation with trace norm regularized objective. Both of them show that recovery guarantees could be attained with lower sample complexity when features are perfect. However, if features are imperfect, such models cannot recover the underlying matrix and could suffer poor performance in practice. We will have a detailed discussion on inductive matrix completion model in Section 3.

Our proposed model is also related to the family of dirty statistical models [36], where the model parameter is expressed as the sum of a number of parameter components, each of which has its own structure. Dirty statistical models have been proposed mostly for robust matrix completion, graphical model estimation, and multi-task learning to decompose the sparse component (noise) and low-rank component (model parameters) [6, 8, 19]. Our proposed algorithm is completely different. We aim to decompose the model into two parts: the part that can be described by side information and the part that has to be recovered purely by observations.

3 A Dirty Statistical Model for Matrix Completion with Features

Let $R \in \mathbb{R}^{n_1 \times n_2}$ be the underlying rank-k matrix that aims to be recovered, where $k \ll \min(n_1, n_2)$ so that R is low-rank. Let Ω be the set of observed entries sampled from R with cardinality $|\Omega| = m$. Furthermore, let $X \in \mathbb{R}^{n_1 \times d_1}$ and $Y \in \mathbb{R}^{n_2 \times d_2}$ be the feature set, where each row $\mathbf{x}_i$ (or $\mathbf{y}_i$) denotes the feature of the i-th row (or column) entity of R. Both $d_1, d_2 \leq \min(n_1, n_2)$ but can be either smaller or larger than k. Thus, given a set of observations Ω and the feature set X and Y as side information, the goal is to recover the underlying low rank matrix R.

To begin with, consider an ideal case where the given features are "perfect" in the following sense:

$$\text{col}(R) \subseteq \text{col}(X) \ \text{ and } \ \text{row}(R) \subseteq \text{col}(Y). \tag{1}$$

Such a feature set can be thought as perfect since it fully describes the true latent feature space of R. Then, instead of recovering the low rank matrix R directly, one can recover a smaller matrix

$M \in \mathbb{R}^{d_1 \times d_2}$ such that $R = XMY^T$. The resulting formulation, called inductive matrix completion (or IMC in brief) [18], is shown to be both theoretically preferred [18, 35] and useful in real applications [37, 29]. Details of this model can be found in [18, 35].

However, in practice, most given features X and Y will not be perfect. In fact, they could be quite noisy or only weakly correlated to the latent feature space of R. Though in some cases applying IMC with imperfect X, Y might still yield decent performance, in many other cases, the performance drastically drops when features become noisy. This weakness of IMC can also be empirically seen in Section 5. Therefore, a more robust model is desired to better handle noisy features.

We now introduce a dirty statistical model for matrix completion with (possibly noisy) features. The core concept of our model is to learn the underlying matrix by balancing feature information and observations. Specifically, we propose to learn R jointly from two parts, one is the low rank estimate from feature space XMY^T, and the other part N is the part outside the feature space. Thus, N can be used to capture the information that noisy features fail to describe, which is then estimated by pure observations. Naturally, both XMY^T and N are preferred to be low rank since they are aggregated to estimate a low rank matrix R. This further leads a preference on M to be low rank as well, since one could expect only a small subspace of X and a subspace of Y are jointly effective to form the low rank space XMY^T. Putting all of above together, we consider to solve the following problem:

$$\min_{M,N} \sum_{(i,j)\in\Omega} \ell((XMY^T + N)_{ij}, R_{ij}) + \lambda_M \|M\|_* + \lambda_N \|N\|_*, \qquad (2)$$

where M and N are regularized with trace norm because of the low rank prior. The underlying matrix R can thus be estimated by $XM^*Y^T + N^*$. We refer our model as DirtyIMC for convenience.

To solve the convex problem (2), we propose an alternative minimization scheme to solve N and M iteratively. Our algorithm is stated in details in Appendix A. One remark of this algorithm is that it is guaranteed to converge to a global optimal, since the problem is jointly convex with M and N.

The parameters λ_M and λ_N are crucial for controlling the importance between features and residual. When $\lambda_M = \infty$, M will be enforced to 0, so features are disregarded and (2) becomes a standard matrix completion objective. Another special case is $\lambda_N = \infty$, in which N will be enforced to 0 and the objective becomes IMC. Intuitively, with an appropriate ratio λ_M/λ_N, the proposed model can incorporate useful part of features, yet be robust to noisy part by compensating from pure observations. Some natural questions arise from here: How to quantify the quality of features? What is the right λ_M and λ_N given a feature set? And beyond intuition, how much can we benefit from features using our model *in theory*? We will formally answer these questions in Section 4.

4 Theoretical Analysis

Now we analyze the usefulness of features in our model under a theoretical perspective. We first quantify the quality of features and show that with reasonably good features, our model achieves recovery with lower sample complexity. Finally, we compare our results to matrix completion and IMC. Due to space limitations, detailed proofs of Theorems and Lemmas are left in Appendix B.

4.1 Preliminaries

Recall that our goal is to recover a rank-k matrix R given observed entry set Ω, feature set X and Y described in Section 3. To recover the matrix with our model (Equation (2)), it is equivalent to solve the hard-constraint problem:

$$\min_{M,N} \sum_{(i,j)\in\Omega} \ell((XMY^T + N)_{ij}, R_{ij}), \quad \text{subject to } \|M\|_* \le \mathcal{M}, \|N\|_* \le \mathcal{N}. \qquad (3)$$

For simplicity, we will consider $d = \max(d_1, d_2) = O(1)$ so that feature dimensions do not grow as a function of n. We assume each entry $(i,j) \in \Omega$ is sampled i.i.d. under an unknown distribution with index set $\{(i_\alpha, j_\alpha)\}_{\alpha=1}^m$. Also, each entry of R is assumed to be upper bounded, i.e. $\max_{ij} |R_{ij}| \le \mathcal{R}$ (so that trace norm of R is in $O(\sqrt{n_1 n_2})$). Such circumstance is consistent with real scenarios like the Netflix problem where users can rate movies with scale from 1 to 5. For convenience, let $\theta = (M, N)$ be any feasible solution, and $\Theta = \{(M,N) \mid \|M\|_* \le \mathcal{M}, \|N\|_* \le \mathcal{N}\}$ be the feasible solution set. Also, let $f_\theta(i,j) = \mathbf{x}_i^T M \mathbf{y}_j + N_{ij}$ be the estimation function for R_{ij} parameterized by θ, and $F_\Theta = \{f_\theta \mid \theta \in \Theta\}$ be the set of feasible functions. We are interested in the following two "ℓ-risk" quantities:

- Expected ℓ-risk: $R_\ell(f) = \mathbb{E}_{(i,j)}\big[\ell(f(i,j), R_{ij})\big]$.

- Empirical ℓ-risk: $\hat{R}_\ell(f) = \frac{1}{m}\sum_{(i,j)\in\Omega} \ell(f(i,j), R_{ij})$.

Thus, our model is to solve for θ^* that parameterizes $f^* = \arg\min_{f\in F_\Theta} \hat{R}_\ell(f)$, and it is sufficient to show that recovery can be attained if $R_\ell(f^*)$ approaches to zero with large enough n and m.

4.2 Measuring the Quality of Features

We now link the quality of features to Rademacher complexity, a learning theoretic tool to measure the complexity of a function class. We will show that quality features result in a lower model complexity and thus a smaller error bound. Under such a viewpoint, the upper bound of Rademacher complexity could be used for measuring the quality of features.

To begin with, we apply the following Lemma to bound the expected ℓ-risk.

Lemma 1 (Bound on Expected ℓ-risk [2])**.** *Let ℓ be a loss function with Lipschitz constant L_ℓ bounded by $\mathcal{B}$ with respect to its first argument, and δ be a constant where $0 < \delta < 1$. Let $\mathfrak{R}(F_\Theta)$ be the Rademacher complexity of the function class F_Θ (w.r.t. Ω and associated with ℓ) defined as:*

$$\mathfrak{R}(F_\Theta) = \mathbb{E}_\sigma\big[\sup_{f\in F_\Theta} \frac{1}{m}\sum_{\alpha=1}^{m} \sigma_\alpha \ell(f(i_\alpha, j_\alpha), R_{i_\alpha j_\alpha})\big], \tag{4}$$

where each σ_α takes values $\{\pm 1\}$ with equal probability. Then with probability at least $1-\delta$, for all $f \in F_\Theta$ we have:

$$R_\ell(f) \leq \hat{R}_\ell(f) + 2\mathbb{E}_\Omega\big[\mathfrak{R}(F_\Theta)\big] + \mathcal{B}\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Apparently, to guarantee a small enough R_ℓ, both $\hat{R}_\ell$ and model complexity $\mathbb{E}_\Omega\big[\mathfrak{R}(F_\Theta)\big]$ have to be bounded. The next key lemma shows that, the model complexity term $\mathbb{E}_\Omega\big[\mathfrak{R}(F_\Theta)\big]$ is related to the feature quality in matrix completion context.

Before diving into the details, we first provide an intuition on the meaning of "good" features. Consider any imperfect feature set which violates (1). One can imagine such feature set is perturbed by some misleading noise which is not correlated to the true latent features. However, features should still be effective if such noise does not weaken the true latent feature information too much. Thus, if a large portion of true latent features lies on the informative part of the feature spaces X and Y, they should still be somewhat informative and helpful for recovering the matrix R.

More formally, the model complexity can be bounded in terms of $\mathcal{M}$ and $\mathcal{N}$ by the following lemma:

Lemma 2. *Let $\mathcal{X} = \max_i \|\mathbf{x}_i\|_2$, $\mathcal{Y} = \max_i \|\mathbf{y}_i\|_2$ and $n = \max(n_1, n_2)$. Then the model complexity of function class F_Θ is upper bounded by:*

$$\mathbb{E}_\Omega\big[\mathfrak{R}(F_\Theta)\big] \leq 2L_\ell\mathcal{M}\mathcal{X}\mathcal{Y}\sqrt{\frac{\log 2d}{m}} + \min\left\{2L_\ell\mathcal{N}\sqrt{\frac{\log 2n}{m}}, \sqrt{9CL_\ell\mathcal{B}\frac{\mathcal{N}(\sqrt{n_1}+\sqrt{n_2})}{m}}\right\}.$$

Then, by Lemma 1 and 2, one could carefully construct a feasible solution set (by setting $\mathcal{M}$ and $\mathcal{N}$) such that both $\hat{R}_\ell(f^*)$ and $\mathbb{E}_\Omega\big[\mathfrak{R}(F_\Theta)\big]$ are controlled to be reasonably small. We now suggest a witness pair of $\mathcal{M}$ and $\mathcal{N}$ constructed as follows. Let γ be defined as:

$$\gamma = \min\left(\frac{\min_i \|\mathbf{x}_i\|}{\mathcal{X}}, \frac{\min_i \|\mathbf{y}_i\|}{\mathcal{Y}}\right).$$

Let $\mathcal{T}_\mu(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ be the thresholding operator where $\mathcal{T}_\mu(x) = x$ if $x \geq \mu$ and $\mathcal{T}_\mu(x) = 0$ otherwise. In addition, let $X = \sum_{i=1}^{d_1} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ be the reduced SVD of X, and define $X_\mu = \sum_{i=1}^{d_1} \sigma_1 \mathcal{T}_\mu(\sigma_i/\sigma_1)\mathbf{u}_i\mathbf{v}_i^T$ to be the "μ-informative" part of X. The ν-informative part of Y, denoted as Y_ν, can also be defined similarly. Now consider setting $\mathcal{M} = \|\hat{M}\|_*$ and $\mathcal{N} = \|R - X_\mu \hat{M} Y_\nu^T\|_*$, where

$$\hat{M} = \arg\min_M \|X_\mu M Y_\nu^T - R\|_F^2 = (X_\mu^T X_\mu)^{-1} X_\mu^T R Y_\nu (Y_\nu^T Y_\nu)^{-1}$$

is the optimal solution for approximating R under the informative feature space X_μ and Y_ν. Then the following lemma shows that the trace norm of $\hat{M}$ will not grow as n increases.

Lemma 3. *Fix $\mu, \nu \in (0, 1]$, and let $\hat{d} = \min(\mathit{rank}(X_\mu), \mathit{rank}(Y_\nu))$. Then with some universal constant C':*

$$\|\hat{M}\|_* \leq \frac{\hat{d}}{C'\mu^2\nu^2\gamma^2\mathcal{X}\mathcal{Y}}.$$

Moreover, by combining Lemma 1 - 3, we can upper bound $R_\ell(f^*)$ of DirtyIMC as follows:

Theorem 1. *Consider problem* (3) *with $\mathcal{M} = \|\hat{M}\|_*$ and $\mathcal{N} = \|R - X_\mu \hat{M} Y_\nu^T\|_*$. Then with probability at least $1-\delta$, the expected ℓ-risk of an optimal solution (N^*, M^*) will be bounded by:*

$$R_\ell(f^*) \le \min\left\{4L_\ell\mathcal{N}\sqrt{\frac{\log 2n}{m}}, \sqrt{36CL_\ell\mathcal{B}\frac{\mathcal{N}(\sqrt{n_1}+\sqrt{n_2})}{m}}\right\} + \frac{4L_\ell\hat{d}}{C'\mu^2\nu^2\gamma^2}\sqrt{\frac{\log 2d}{m}} + \mathcal{B}\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

4.3 Sample Complexity Analysis

From Theorem 1, we can derive the following sample complexity guarantee of our model. For simplicity, we assume $k = O(1)$ so it will not grow as n increases in the following discussion.

Corollary 1. *Suppose we aim to "ϵ-recover" R where $\mathbb{E}_{(i,j)}\left[\ell(N_{ij} + XMY_{ij}^T, R_{ij})\right] < \epsilon$ given an arbitrarily small ϵ. Then for DirtyIMC model, $O(\min(\mathcal{N}\sqrt{n}, \mathcal{N}^2\log n)/\epsilon^2)$ observations are sufficient for ϵ-recovery provided a sufficiently large n.*

Corollary 1 suggests that the sample complexity of our model only depends on the trace norm of residual $\mathcal{N}$. This matches the intuition of good features stated in Section 4.2 because $X\hat{M}Y^T$ will cover most part of R if features are good, and as a result, $\mathcal{N}$ will be small and one can enjoy small sample complexity by exploiting quality features.

We also compare our sample complexity result with other models. First, suppose features are perfect (so that $\mathcal{N} = O(1)$), our result suggests that only $O(\log n)$ samples are required for recovery. This matches the result of [35], in which the authors show that given perfect features, $O(\log n)$ observations are enough for exact recovery by solving the IMC objective. However, IMC does not guarantee recovery when features are not perfect, while our result shows that recovery is still attainable by DirtyIMC with $O(\min(\mathcal{N}\sqrt{n}, \mathcal{N}^2\log n)/\epsilon^2)$ samples. We will also empirically justify this result in Section 5.

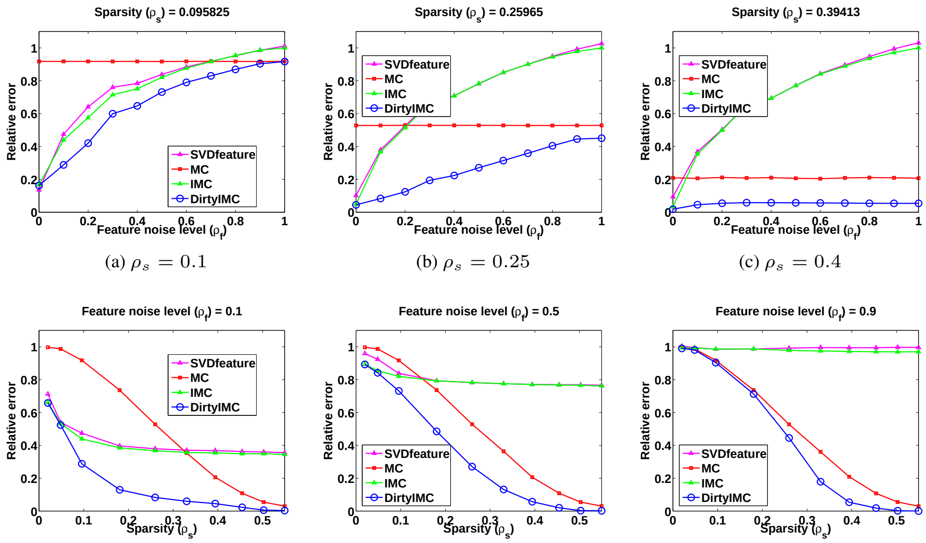
On the other hand, for standard matrix completion (i.e. no features are considered), the most well-known guarantee is that under certain conditions, one can achieve $O(n \operatorname{poly}\log n)$ sample complexity for both ϵ-recovery [34] and exact recovery [5]. However, these bounds only hold with distributional assumptions on observed entries. For sample complexity without any distributional assumptions, Shamir et al. [32] recently showed that $O(n^{3/2})$ entries are sufficient for ϵ-recovery, and this bound is tight if no further distribution of observed entries is assumed. Compared to those results, our analysis also requires no assumptions on distribution of observed entries, and our sample complexity yields $O(n^{3/2})$ as well in the worst case, by the fact that $\mathcal{N} \le \|R\|_* = O(n)$. Notice that it is reasonable to meet the lower bound $\Omega(n^{3/2})$ even given features, since in an extreme case, X, Y could be random matrices and have no correlation to R, and thus the given information is as same as that in standard matrix completion.

However, in many applications, features will be far from random, and our result provides a theoretical insight to show that features can be useful even if they are imperfect. Indeed, as long as features are informative enough such that $\mathcal{N} = o(n)$, our sample complexity will be asymptotically lower than $O(n^{3/2})$. Here we provide two concrete instances for such a scenario. In the first scenario, we consider the rank-k matrix R to be generated from random orthogonal model [5] as follows:

Theorem 2. *Let $R \in \mathbb{R}^{n\times n}$ be generated from random orthogonal model, where $U = \{\mathbf{u}_i\}_{i=1}^k$, $V = \{\mathbf{v}_i\}_{i=1}^k$ are random orthogonal bases, and $\sigma_1 \ldots \sigma_k$ are singular values with arbitrary magnitude. Let σ_t be the largest singular value such that $\lim_{n\to\infty}\sigma_t/\sqrt{n} = 0$. Then, given the noisy features X, Y where $X_{:i} = \mathbf{u}_i$ (and $Y_{:i} = \mathbf{v}_i$) if $i < t$ and $X_{:i}$ (and $V_{:i}$) be any basis orthogonal to U (and V) if $i \ge t$, $o(n)$ samples are sufficient for DirtyIMC to achieve ϵ-recovery.*

Theorem 2 suggests that, under random orthogonal model, if features are not too noisy in the sense that noise only corrupts the true subspace associated with smaller singular values, we can approximately recover R with only $o(n)$ observations. An empirical justification for this result is presented in Appendix C. Another scenario is to consider R to be the product of two rank-k Gaussian matrices:

Theorem 3. *Let $R = UV^T$ be a rank-k matrix, where $U, V \in \mathbb{R}^{n\times k}$ are true latent row/column features with each $U_{ij}, V_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Suppose now we are given a feature set X, Y where $g(n)$ row items and $h(n)$ column items have corrupted features. Moreover, each corrupted row/column item has perturbed feature $\mathbf{x}_i = \mathbf{u}_i + \Delta\mathbf{u}_i$ and $\mathbf{y}_i = \mathbf{v}_i + \Delta\mathbf{v}_i$, where $\|\Delta\mathbf{u}\|_\infty \le \xi_1$ and*



(a) $\rho_s = 0.1$

(b) $\rho_s = 0.25$

(c) $\rho_s = 0.4$

(d) $\rho_f = 0.1$

(e) $\rho_f = 0.5$

(f) $\rho_f = 0.9$

Figure 1: Performance of various methods for matrix completion under different sparsity and feature quality. Compared to other feature-based completion methods, the top figures show that DirtyIMC is less sensitive to noisy features with each ρ_s, and the bottom figures show that error of DirtyIMC always decreases to 0 with more observations given any feature quality.

$\|\Delta\mathbf{v}\|_\infty \leq \xi_2$ *with some constants* ξ_1 *and* ξ_2. *Then for DirtyIMC model* (3), *with high probability,* $O\big(\max(\sqrt{g(n)}, \sqrt{h(n)})n \log n\big)$ *observations are sufficient for* ϵ*-recovery.*

Theorem 3 suggests that, if features have good quality in the sense that items with corrupted features are not too many, for example $g(n), h(n) = O(\log n)$, then sample complexity of DirtyIMC can be $O(n \log n \sqrt{\log n}) = o(n^{3/2})$ as well. Thus, both Theorem 2 and 3 provide concrete examples showing that given imperfect yet informative features, the sample complexity of our model can be asymptotically lower than the lower bound of pure matrix completion (which is $\Omega(n^{3/2})$).

5 Experimental Results

In this section, we show the effectiveness of the DirtyIMC model (2) for matrix completion with features on both synthetic datasets and real-world applications. For synthetic datasets, we show that DirtyIMC model better recovers low rank matrices under various quality of features. For real applications, we consider relationship prediction and semi-supervised clustering, where the current state-of-the-art methods are based on matrix completion and IMC respectively. We show that by applying DirtyIMC model to these two problems, we can further improve performance by making better use of features.

5.1 Synthetic Experiments

We consider matrix recovery with features on synthetic data generated as follows. We create a low rank matrix $R = UV^T$, as the true latent row/column space $U, V \in \mathbb{R}^{200\times 20}$, $U_{ij}, V_{ij} \sim \mathcal{N}(0, 1/20)$. We then randomly sample ρ_s percent of entries Ω from R as observations, and construct a perfect feature set $X^*, Y^* \in \mathbb{R}^{200\times 40}$ which satisfies (1). To examine performance under different quality of features, we generate features X, Y with a noise parameter ρ_f, where X and Y will be derived by replacing ρ_f percent of bases of X^* (and Y^*) with bases orthogonal to X^* (and Y^*). We then consider recovering the underlying matrix R given X, Y and a subset Ω of R.

We compare our DirtyIMC model (2) with standard trace-norm regularized matrix completion (MC) and two other feature-based completion methods: IMC [18] and SVDfeature [9]. The standard relative error $\|\hat{R} - R\|_F / \|R\|_F$ is used to evaluate a recovered matrix $\hat{R}$. For each method, we select parameters from the set $\{10^\alpha\}_{\alpha=-3}^{2}$ and report the one with the best recovery. All results are averaged over 5 random trials.

Figure 1 shows the recovery of each method under each sparsity level $\rho_s = 0.1, 0.25, 0.4$, and each feature noise level $\rho_f = 0.1$, 0.5 and 0.9. We first observe that in the top figures, IMC and

Method	DirtyIMC	MF-ALS [16]	IMC [18]	HOC-3	HOC-5 [12]
Accuracy	**0.9474**±0.0009	0.9412±0.0011	0.9139±0.0016	0.9242±0.0010	0.9297±0.0011
AUC	**0.9506**	0.9020	0.9109	0.9432	0.9480

Table 1: Relationship prediction on Epinions. Compared with other approaches, DirtyIMC model gives the best performance in terms of both accuracy and AUC.

SVDfeature perform similarly under different ρ_s. This suggests that with sufficient observations, performance of IMC and SVDfeature mainly depend on feature quality and will not be affected much by the number of observations. As a result, given good features (1d), they achieve smaller error compared to MC with few observations, but as features become noisy (1e-1f), they suffer poor performance by trying to learn the underlying matrix under biased feature spaces. Another interesting finding is that when good features are given (1d), IMC (and SVDfeature) still fails to achieve 0 relative error as the number of observations increases, which reconfirms that IMC cannot guarantee recoverability when features are not perfect. On the other hand, we see that performance of DirtyIMC can be improved by both better features or more observations. In particular, it makes use of informative features to achieve lower error compared to MC and is also less sensitive to noisy features compared to IMC and SVDfeature. Some finer recovery results on ρ_s and ρ_f can be found in Appendix C.

5.2 Real-world Applications

Relationship Prediction in Signed Networks. As the first application, we consider relationship prediction problem in an online review website Epinions [26], where people can write reviews and trust or distrust others based on their reviews. Such social network can be modeled as a signed network where trust/distrust are modeled as positive/negative edges between entities [24], and the problem is to predict unknown relationship between any two users given the network. A state-of-the-art approach is the low rank model [16, 12] where one can first conduct matrix completion on adjacency matrix and then use the sign of completed matrix for relationship prediction. Therefore, if features of users are available, we can also consider low rank model by using our model for matrix completion step. This approach can be regarded as an improvement over [16] by incorporating feature information.

In this dataset, there are about $n = 105$K users and $m = 807$K observed relationship pairs where 15% relationships are distrust. In addition to who-trust-to-whom information, we also have user feature matrix $Z \in \mathbb{R}^{n \times 41}$ where for each user a 41-dimensional feature is collected based on the user's review history, such as number of positive/negative reviews the user gave/received. We then consider the low-rank model in [16] where matrix completion is conducted by DirtyIMC with non-convex relaxation (5) (DirtyIMC), IMC [18] (IMC), and matrix factorization proposed in [16] (MF-ALS), along with another two prediction methods, HOC-3 and HOC-5 [12]. Note that both row and column entities are users so $X = Y = Z$ is set for both DirtyIMC and IMC model.

We conduct the experiment using 10-fold cross validation on observed edges, where the parameters are chosen from the set $\sqcup_{\alpha=-3}^{2}\{10^\alpha, 5 \times 10^\alpha\}$. The averaged accuracy and AUC of each method are reported in Table 1. We first observe that IMC performs worse than MF-ALS even though IMC takes features into account. This is because features are only weakly related to relationship matrix, and as a result, IMC is misled by such noisy features. On the other hand, DirtyIMC performs the best among all prediction methods. In particular, it performs slightly better than MF-ALS in terms of accuracy, and much better in terms of AUC. This shows DirtyIMC can still exploit weakly informative features without being trapped by noisy features.

Semi-supervised Clustering. We now consider semi-supervised clustering problem as another application. Given n items, the item feature matrix $Z \in \mathbb{R}^{n \times d}$, and m pairwise constraints specifying whether item i and j are similar or dissimilar, the goal is to find a clustering of items such that most similar items are within the same cluster.

We notice that the problem can indeed be solved by matrix completion. Consider $S \in \mathbb{R}^{n \times n}$ to be the signed similarity matrix defined as $S_{ij} = 1$ (or -1) if item i and j are similar (or dissimilar), and 0 if similarity is unknown. Then solving semi-supervised clustering becomes equivalent to finding a clustering of the symmetric signed graph S, where the goal is to cluster nodes so that most edges within the same group are positive and most edges between groups are negative [12]. As a result, a matrix completion approach [12] can be applied to solve the signed graph clustering problem on S.

Apparently, the above solution is not optimal for semi-supervised clustering as it disregards features. Many semi-supervised clustering algorithms are thus proposed by taking both item features

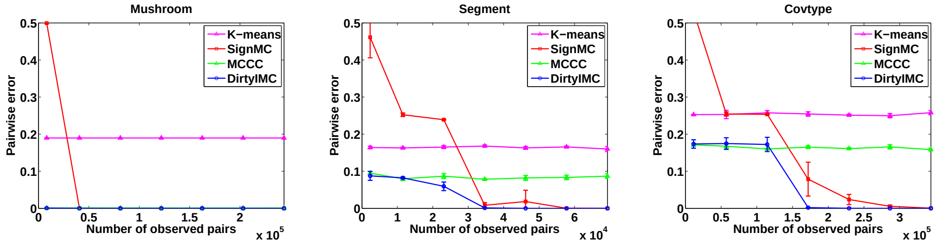


Figure 2: Semi-supervised clustering on real-world datasets. For Mushroom dataset where features are almost ideal, both MCCC and DirtyIMC achieve 0 error rate. For Segment and Covtype where features are more noisy, our model outperforms MCCC as its error decreases given more constraints.

	number of items n	feature dimension d	number of clusters k
Mushrooms	8124	112	2
Segment	2319	19	7
Covtype	11455	54	7

Table 2: Statistics of semi-supervised clustering datasets.

and constraints into consideration [13, 25, 37]. The current state-of-the-art method is the MCCC algorithm [37], which essentially solves semi-supervised clustering with IMC objective. In [37], the authors show that by running k-means on the top-k eigenvectors of the completed matrix ZMZ^T, MCCC outperforms other state-of-the-art algorithms [37].

We now consider solving semi-supervised clustering with our DirtyIMC model. Our algorithm, summarized in Algorithm 2 in Appendix D, first completes the pairwise matrix with DirtyIMC objective (2) instead of IMC (with both X, Y are set as Z), and then runs k-means on the top-k eigenvectors of the completed matrix to obtain a clustering. This algorithm can be viewed as an improved version of MCCC to handle noisy features Z.

We now compare our algorithm with k-means, signed graph clustering with matrix completion [12] (SignMC) and MCCC [37]. Note that since MCCC has been shown to outperform most other state-of-the-art semi-supervised clustering algorithms in [37], comparing with MCCC is sufficient to demonstrate the effectiveness of our algorithm. We perform each method on three real-world datasets: Mushrooms, Segment and Covtype [1]. All of them are classification benchmarks where features and ground-truth class of items are both available, and their statistics are summarized in Table 2. For each dataset, we randomly sample $m = [1, 5, 10, 15, 20, 25, 30] \times n$ pairwise constraints, and perform each algorithm to derive a clustering π, where π_i is the cluster index of item i. We then evaluate π by the following pairwise error to ground-truth:

$$\frac{n(n-1)}{2}\left(\sum_{(i,j):\pi_i^*=\pi_j^*} \mathbf{1}(\pi_i \neq \pi_j) + \sum_{(i,j):\pi_i^*\neq\pi_j^*} \mathbf{1}(\pi_i = \pi_j)\right)$$

where π_i^* is the ground-truth class of item i.

Figure 2 shows the result of each method on all three datasets. We first see that for Mushrooms dataset where features are perfect (100% training accuracy can be attained by linear-SVM for classification), both MCCC and DirtyIMC can obtain a perfect clustering, which shows that MCCC is indeed effective with perfect features. For Segment and Covtype datasets, we observe that the performance of k-means and MCCC are dominated by feature quality. Although MCCC still benefits from constraint information as it outperforms k-means, it clearly does not make the best use of constraints, as its performance does not improves even if number of constraints increases. On the other hand, the error rate of SignMC can always decrease down to 0 by increasing m. However, since it disregards features, it suffers from a much higher error rate than methods with features when constraints are few. We again see DirtyIMC combines advantage from MCCC and SignMC, as it makes use of features when few constraints are observed yet leverages constraint information simultaneously to avoid being trapped by feature noise. This experiment shows that our model outperforms state-of-the-art approaches for semi-supervised clustering.

Acknowledgement. We thank David Inouye and Hsiang-Fu Yu for helpful comments and discussions. This research was supported by NSF grants CCF-1320746 and CCF-1117055.

[1] All datasets are available at `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. For Covtype, we subsample from the entire dataset to make each cluster has balanced size.

References

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *JMLR*, 10:803–826, 2009.

[2] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2003.

[3] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA 02178-9998, 1999.

[4] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[5] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011.

[7] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, 2010.

[8] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 2012.

[9] T. Chen, W. Zhang, Q. Lu, K. Chen, Z. Zheng, and Y. Yu. SVDFeature: A toolkit for feature-based collaborative filtering. *JMLR*, 13:3619–3622, 2012.

[10] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. In *ICML*, 2014.

[11] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *JMLR*, 15(1):2213–2238, 2014.

[12] K.-Y. Chiang, C.-J. Hsieh, N. Natarajan, I. S. Dhillon, and A. Tewari. Prediction and clustering in signed networks: A local to global perspective. *JMLR*, 15:1177–1213, 2014.

[13] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[14] U. Feige and G. Schechtman. On the optimality of the random hyperplane rounding technique for max cut. *Random Struct. Algorithms*, 20(3):403–440, 2002.

[15] L. Grippo and M. Sciandrone. Globally convergent block-coordinate techniques for unconstrained optimization. *Optimization Methods and Software*, 10:587–637, 1999.

[16] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon. Low rank modeling of signed networks. In *KDD*, 2012.

[17] C.-J. Hsieh and P. A. Olsan. Nuclear norm minimization via active subspace selection. In *ICML*, 2014.

[18] P. Jain and I. S. Dhillon. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013.

[19] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, 2010.

[20] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, pages 793 – 800, 2008.

[21] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *JMLR*, 2010.

[22] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42:30–37, 2009.

[23] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[24] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, 2010.

[25] Z. Li and J. Liu. Constrained clustering by spectral kernel learning. In *ICCV*, 2009.

[26] P. Massa and P. Avesani. Trust-aware bootstrapping of recommender systems. In *Proceedings of ECAI 2006 Workshop on Recommender Systems*, pages 29–33, 2006.

[27] R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *JMLR*, 2003.

[28] A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *KDD*, pages 141–149, 2011.

[29] N. Natarajan and I. S. Dhillon. Inductive matrix completion for predicting gene-disease associations. *Bioinformatics*, 30(12):60–68, 2014.

[30] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *JMLR*, 13(1):1665–1697, 2012.

[31] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math*, pages 1707–1739, 2009.

[32] O. Shamir and S. Shalev-Shwartz. Matrix completion with the trace norm: Learning, bounding, and transducing. *JMLR*, 15(1):3401–3423, 2014.

[33] D. Shin, S. Cetintas, K.-C. Lee, and I. S. Dhillon. Tumblr blog recommendation with boosted inductive matrix completion. In *CIKM*, pages 203–212, 2015.

[34] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *COLT*, pages 545–560, 2005.

[35] M. Xu, R. Jin, and Z.-H. Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *NIPS*, 2013.

[36] E. Yang and P. Ravikumar. Dirty statistical models. In *NIPS*, 2013.

[37] J. Yi, L. Zhang, R. Jin, Q. Qian, and A. Jain. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *ICML*, 2013.

[38] K. Zhong, P. Jain, and I. S. Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *International Conference on Algorithmic Learning Theory(ALT)*, 2015.

Algorithm 1 Alternative Minimization for DirtyIMC with Squared Loss

```
Input: feature matrix $X, Y$, parameters $(\lambda_M, \lambda_N)$ in objective (2), max iteration $t_{max}$.
$t = 0, M^{(t)} \leftarrow 0, N^{(t)} \leftarrow 0$.
while Not converged and $t < t_{max}$ do
    Solve $M^{(t+1)} \leftarrow \arg\min_M \sum_{(i,j)\in\Omega}(XMY^T_{ij} - (R - N^{(t)})_{ij})^2 + \lambda_M \|M\|_*$
    Solve $N^{(t+1)} \leftarrow \arg\min_N \sum_{(i,j)\in\Omega}(N_{ij} - (R - XM^{(t+1)}Y^T)_{ij})^2 + \lambda_N \|N\|_*$
    $t \leftarrow t + 1$
end while
return recovered matrix $XM^{(t)}Y^T + N^{(t)}$.
```

Appendix A: Solving DirtyIMC Objectives

To solve problem (2), we propose an alternative minimization scheme where at each step we fix one of the variables (M or N) and solve for the other. For simplicity, here we focus on the case where ℓ is squared loss, which is also considered in our experiments. The algorithm is summarized in Algorithm 1. As one variable is fixed, the subproblem reduces to either standard matrix completion or IMC, which is easy to solve as discussed below. This algorithm can be viewed as applying a block coordinate descent algorithm on convex (but non-smooth) function, and thus is guaranteed to converge to global optimal using standard analysis (e.g. [15]).

We now briefly discuss how to solve two subproblems in Algorithm 1. First, when fixing N, the subproblem becomes an IMC objective with observed matrix to be $R - N$. We then apply proximal gradient descent to update M. Notice that in our setting, feature dimensions (d_1, d_2) are much smaller than number of entities (n_1, n_2). Therefore, for small d, it is relatively inexpensive to compute a full SVD for a $d_1 \times d_2$ matrix in each proximal step.

On the other hand, when fixing M, the subproblem becomes standard matrix completion problem for the residual matrix $R - XMY^T$. We then apply active subspace selection algorithm (Active-ALT) [17] to solve the matrix completion problem.

Another possibility is to consider the non-convex relaxation of problem (2) as:

$$\min_{U,V,W,H} \sum_{(i,j)\in\Omega} \ell((XU^TVY^T + W^TH)_{ij}, R_{ij}) + \frac{\lambda_M}{2}(\|W\|_F^2 + \|H\|_F^2) + \frac{\lambda_N}{2}(\|U\|_F^2 + \|V\|_F^2), \quad (5)$$

in which M, N is factorized to low rank matrices $U \in \mathbb{R}^{d_1 \times k_1}, V \in \mathbb{R}^{d_2 \times k_1}$ and $W \in \mathbb{R}^{n_1 \times k_2}, H \in \mathbb{R}^{n_2 \times k_2}$. A similar alternative minimization scheme, i.e. fix three variables and solve for the other, can be applied to obtain a solution for U, V, W, H. Although problem (5) is equivalent to the convex problem (2) if $k_1 \geq \text{rank}(M^*)$ and $k_2 \geq \text{rank}(N^*)$ [34], it is not jointly convex for all variables. So unlike Algorithm 1, using alternative minimization to solve (5) may not obtain the global optimum. However, the analysis in [3] shows that the algorithm converges to stationary points if each subproblem has a unique minimizer, which is indeed the case in (5) because of the regularizations. Researchers found that such non-convex relaxation to be useful since it is easier to solve, and empirically yields a competitive result compared to convex problem [22].

Finally, we notice that a recently proposed method "Boosted IMC" [33] could also be represented as a special case of our alternative scheme for non-convex relaxation (5). The method could be viewed as an one iteration heuristic of Algorithm 1 (i.e. $t_{max} = 1$), in which they first solve $N^{(1)}$ and then solve $M^{(1)}$ using matrix factorization. Although this method is proposed as a heuristic for Blog recommendation rather than an algorithm for solving a formal defined matrix completion objective, it could also be interpreted as an algorithm that approximately solves our DirtyIMC model. We also compare our DirtyIMC with Boosted IMC in Appendix C.

Appendix B: Proofs

Proof of Lemma 2

Proof. To begin with, we introduce a lemma to bound the Rademacher complexity for the function class with bounded trace norm.

Lemma 4. *Let $S_w = \{W \in \mathbb{R}^{n\times n} \mid \|W\|_* \le \mathcal{W}\}$ and $\mathcal{A} = \max_i \|A_i\|_2$, where each $A_i \in \mathbb{R}^{n\times n}$, then:*

$$\mathbb{E}_\sigma\big[\sup_{W\in S_w}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \mathit{trace}(WA_i)\big] \le 2\mathcal{A}\mathcal{W}\sqrt{\frac{\log 2n}{m}}.$$

This Lemma is a special case of Theorem 1 in [20] with the fact that the dual norm of the matrix 2-norm is trace norm. Thus, by using Rademacher contraction principle (e.g. Lemma 5 in [27]), $\mathfrak{R}(F_\Theta)$ can be written as:

$$\begin{aligned}
\mathfrak{R}(F_\Theta) &\le L_\ell \mathbb{E}_\sigma\big[\sup_{\theta\in\Theta}\frac{1}{m}\sum_{\sigma=1}^{m}\sigma_\alpha(XMY^T+N)_{i_\alpha j_\alpha}\big] \\
&= L_\ell \mathbb{E}_\sigma\big[\sup_{\|M\|_*\le\mathcal{M}}\frac{1}{m}\sum_{\sigma=1}^{m}\sigma_\alpha \mathbf{x}_{i_\alpha}^T M\mathbf{y}_{j_\alpha}\big] + L_\ell \mathbb{E}_\sigma\big[\sup_{\|N\|_*\le\mathcal{N}}\frac{1}{m}\sum_{\sigma=1}^{m}\sigma_\alpha N_{i_\alpha j_\alpha}\big] \\
&= L_\ell \mathbb{E}_\sigma\big[\sup_{\|M\|_*\le\mathcal{M}}\frac{1}{m}\sum_{\alpha=1}^{m}\sigma_\alpha \mathrm{trace}(M\mathbf{y}_{j_\alpha}\mathbf{x}_{i_\alpha}^T)\big] + L_\ell \mathbb{E}_\sigma\big[\sup_{\|N\|_*\le\mathcal{N}}\frac{1}{m}\sum_{\alpha=1}^{m}\sigma_\alpha \mathrm{trace}(N\mathbf{e}_{j_\alpha}\mathbf{e}_{i_\alpha}^T)\big] \\
&\le 2L_\ell\left(\mathcal{M}\max_{i,j}\|\mathbf{y}_j\mathbf{x}_i^T\|_2\sqrt{\frac{\log 2d}{m}} + \mathcal{N}\sqrt{\frac{\log 2n}{m}}\right),
\end{aligned}$$

where the last equation is derived by applied Lemma 4. Since $\max_{i,j}\|\mathbf{y}_j\mathbf{x}_i^T\|_2 = \max_j\|\mathbf{y}_j\|_2\max_i\|\mathbf{x}_i\|_2$, we derive an upper bound of $\mathfrak{R}(F_\Theta)$:

$$\mathbb{E}_\Omega\big[\mathfrak{R}(F_\Theta)\big] \le 2L_\ell\mathcal{M}\mathcal{X}\mathcal{Y}\sqrt{\frac{\log 2d}{m}} + 2L_\ell\mathcal{N}\sqrt{\frac{\log 2n}{m}}. \tag{6}$$

However, in some circumstances, the above bound (6) will become too loose for our sample complexity analysis. As a result, we need to deal with these cases by introducing a tighter bound on trace norm of residual (i.e. $\mathcal{N}$). The following bound mainly follows the proof step in [32], which provides a tighter bound on trace-norm regularized function class. To begin with, we can rewrite $\mathfrak{R}(F_\Theta)$ as:

$$\begin{aligned}
\mathfrak{R}(F_\Theta) &= \mathbb{E}_\sigma\big[\sup_{f\in F_\Theta}\frac{1}{m}\sum_{\alpha=1}^{m}\sigma_\alpha\ell(f(i_\alpha,j_\alpha),R_{i_\alpha,j_\alpha}))\big] \\
&= \mathbb{E}_\sigma\big[\sup_{f\in F_\Theta}\frac{1}{m}\sum_{(i,j)}\Gamma_{ij}\ell(f(i,j),R_{ij})\big],
\end{aligned}$$

where $\Gamma \in \mathbb{R}^{n_1\times n_2}$ with each entry $\Gamma_{ij} = \sum_{\alpha: i_\alpha=i, j_\alpha=j}\sigma_\alpha$. Now, using the same trick in [32], we can divide Γ based on the "hit-time" on entry (i,j) of Ω, with some threshold $p > 0$ whose value will be set later. Formally, let $h_{ij} = |\{\alpha : i_\alpha = i, j_\alpha = j\}|$, and let $A, B \in \mathbb{R}^{n_1\times n_2}$ be defined as:

$$A_{ij} = \begin{cases}\Gamma_{ij}, & \text{if } h_{ij} > p,\\ 0, & \text{otherwise.}\end{cases} \qquad B_{ij} = \begin{cases}0, & \text{if } h_{ij} > p,\\ \Gamma_{ij}, & \text{otherwise.}\end{cases} \tag{7}$$

By construction, $\Gamma = A + B$. Therefore, we can separate $\mathfrak{R}(F_\Theta)$ as:

$$\mathfrak{R}(F_\Theta) = \mathbb{E}_\sigma\big[\sup_{f\in F_\Theta}\frac{1}{m}\sum_{(i,j)}A_{ij}\ell(f(i,j),R_{ij})\big] + \mathbb{E}_\sigma\big[\sup_{f\in F_\Theta}\frac{1}{m}\sum_{(i,j)}B_{ij}\ell(f(i,j),R_{ij})\big]. \tag{8}$$

For the first term of (8), by the assumption $|\ell(f(i,j),R_{ij})| \le \mathcal{B}$, it can be upper bounded by:

$$\frac{\mathcal{B}}{m}\mathbb{E}_\sigma\big[\sum_{(i,j)}|A_{ij}|\big] \le \frac{\mathcal{B}}{\sqrt{p}}$$

by using the Lemma 10 in [32]. Now consider the second term of (8). Again, by using Rademacher contraction principle, it can be upper bounded by:

$$\begin{aligned}
&\frac{L_\ell}{m}\mathbb{E}_\sigma\big[\sup_{f\in F_\Theta}\sum_{(i,j)}B_{ij}f(i,j)\big] \\
=&\frac{L_\ell}{m}\mathbb{E}_\sigma\big[\sup_{M:\|M\|_*\le\mathcal{M}}\sum_{(i,j)}B_{ij}\mathbf{x}_i^T M\mathbf{y}_j\big] + \frac{L_\ell}{m}\mathbb{E}_\sigma\big[\sup_{N:\|N\|_*\le\mathcal{N}}\sum_{(i,j)}B_{ij}N_{ij}\big],
\end{aligned} \tag{9}$$

which is separated by feature-covered part and residual part. We first consider the residual part (i.e. the second term of (9)). By applying Hölder's inequality, the second term of (9) is upper bounded by:

$$\frac{L_\ell}{m} \sup_{N:\|N\|_* \leq \mathcal{N}} \|B\|_2 \|N\|_* = \frac{L_\ell \mathcal{N}}{m} \mathbb{E}_\sigma \big[\|B\|_2 \big] \leq \frac{2.2 C L_\ell \mathcal{N} \sqrt{p}(\sqrt{n_1} + \sqrt{n_2})}{m},$$

where the last inequality is derived by applying Lemma 11 in [32]. Now, for the first term of (9), notice that we can upper bound this term by:

$$\begin{aligned} &\frac{L_\ell}{m} \mathbb{E}_\sigma \big[\sup_{M:\|M\|_* \leq \mathcal{M}} \sum_{\alpha=1}^{m} \sigma_\alpha \mathbf{x}_{i_\alpha}^T M \mathbf{y}_{j_\alpha} \big] \\ &= L_\ell \mathbb{E}_\sigma \big[\sup_{\|M\|_* \leq \mathcal{M}} \frac{1}{m} \sum_{\alpha=1}^{m} \sigma_\alpha \text{trace}(M \mathbf{y}_{j_\alpha} \mathbf{x}_{i_\alpha}^T) \big] \\ &\leq 2 L_\ell \mathcal{M} \max_{i,j} \|\mathbf{y}_j \mathbf{x}_i^T\|_2 \sqrt{\frac{\log 2d}{m}} \\ &= 2 L_\ell \mathcal{M} \mathcal{X} \mathcal{Y} \sqrt{\frac{\log 2d}{m}}. \end{aligned}$$

Therefore, putting back all above upper bound to (8), with p chosen to be $m\mathcal{B}/(2.2CL_\ell\mathcal{N}(\sqrt{n_1} + \sqrt{n_2}))$, we can get another bound on $\mathfrak{R}(F_\Theta)$ by:

$$\mathbb{E}_\Omega \big[\mathfrak{R}(F_\Theta) \big] \leq 2 L_\ell \mathcal{M} \mathcal{X} \mathcal{Y} \sqrt{\frac{\log 2d}{m}} + \sqrt{9 C L_\ell \mathcal{B} \frac{\mathcal{N}(\sqrt{n_1} + \sqrt{n_2})}{m}}. \tag{10}$$

The Theorem thus follows by combining two bounds from (6) and (10). □

Proof of Lemma 3

We first need the following lemma to bound the largest singular value σ_x of feature matrix X (and also σ_y of Y) .

Lemma 5. *Let $X \in \mathbb{R}^{n \times d}$ be a feature matrix. Then there exists a constant C'' (i.e. not a function of n), such that:*

$$\sigma_x \geq C'' \gamma \mathcal{X} \sqrt{n}.$$

Proof. Let $\tilde{\mathbf{x}}_i$ be normalized feature vectors that $\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$ for all $i = 1 \dots n$, so that each $\tilde{\mathbf{x}}_i$ lies on the d dimensional unit sphere $S_d = \{\tilde{\mathbf{x}} \in \mathbb{R}^d \mid \|\tilde{\mathbf{x}}\| = 1\}$. From Lemma 21 of [14], for any $\eta > 0$, the d dimensional unit sphere can be partitioned into $N = (c/\eta)^d$ equal volume cells (denoted as $P_1 \dots P_N$) whose diameter is at most η, where c is some constant. Therefore, if two unit vectors $\mathbf{x}, \mathbf{y}$ are in the same cell P_i, since $\|\mathbf{x} - \mathbf{y}\| \leq \eta$, the angle θ between $\mathbf{x}$ and $\mathbf{y}$ will satisfy

$$\frac{\theta}{2} \leq \sin^{-1}(\frac{\eta}{2\|\mathbf{x}\|}) = \sin^{-1} \frac{\eta}{2},$$

which leads the inner product of $\mathbf{x}$ and $\mathbf{y}$ to be:

$$\mathbf{x}^T \mathbf{y} = \cos\theta = 1 - 2\sin^2(\frac{\theta}{2}) \geq 1 - 2(\frac{\eta}{2})^2 = 1 - \frac{\eta^2}{2}.$$

Thus, taking $\eta = 1$, we can partition the unit sphere into $N = c^d$ cells such that

$$\mathbf{x}^T \mathbf{y} \geq \frac{1}{2}, \text{ if } \mathbf{x}, \mathbf{y} \in P_i.$$

Now reconsider n normalized feature vectors $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n$, each of which belongs to one of the cell P_i. By Pigeonhole Theorem, there exists one cell $P^* \in \{P_i\}_{i=1}^N$ such that at least n/N vectors lie in P^*. Consider any unit vector $\mathbf{w}$ in P^*, then we have $\tilde{\mathbf{x}}_i^T\mathbf{w} \geq \frac{1}{2}$ for all $\tilde{\mathbf{x}}_i \in P^*$. Therefore,

$$\begin{aligned}\|X\mathbf{w}\|_2 &\geq \sqrt{\sum_{i:\mathbf{x}_i\in P^*}(\mathbf{x}_i^T\mathbf{w})^2}\\ &\geq \sqrt{\sum_{i:\mathbf{x}_i\in P^*}\gamma^2\mathcal{X}^2(\tilde{\mathbf{x}}_i^T\mathbf{w})^2}\\ &\geq \gamma\mathcal{X}\sqrt{\frac{n}{N}(\frac{1}{2})^2}\\ &= \left(\frac{1}{2\sqrt{N}}\right)\gamma\mathcal{X}\sqrt{n},\end{aligned}$$

which concludes that

$$\sigma_x \geq C''\gamma\mathcal{X}\sqrt{n}$$

where $C'' = \frac{1}{2\sqrt{N}}$ is a constant with respect to n. □

With Lemma 5, we can now prove the Lemma 3 as follows:

Proof. To begin with, we have:

$$\|X_\mu^T R Y_\nu\|_2 \leq \|X_\mu\|_2\|R\|_2\|Y_\nu\|_2 \leq \sigma_x\sigma_y\|R\|_*.$$

On the other hand, by the closed form solution of $\hat{M}$, we have:

$$\begin{aligned}\|\hat{M}\|_* &\leq \|\hat{M}\|_2\hat{d}\\ &= \|(X_\mu^T X_\mu)^{-1}X_\mu^T R Y(Y_\nu^T Y_\nu)^{-1}\|_2\hat{d}\\ &\leq \frac{\sigma_x\sigma_y\|R\|_*\hat{d}}{\sigma_{xm}^2\sigma_{ym}^2},\end{aligned}$$

where σ_{xm}, σ_{ym} are the smallest singular value of X_μ, Y_ν respectively. Also, by construction of X_μ and Y_ν, we have $\sigma_{xm} \geq \mu\sigma_x$ and $\sigma_{ym} \geq \nu\sigma_y$. Combining Lemma 5, we have:

$$\begin{aligned}\|\hat{M}\|_* &\leq \frac{\|R\|_*\hat{d}}{\mu^2\nu^2\sigma_x\sigma_y}\\ &\leq \frac{\|R\|_*\hat{d}}{C'\sqrt{n_1n_2}\gamma^2\mu^2\nu^2\mathcal{X}\mathcal{Y}},\end{aligned}$$

where C' is a constant independent to n_1, n_2. By the fact that $\|R\|_* \leq \mathcal{R}\sqrt{n_1n_2}$, the lemma is proved. □

Proof of Theorem 2

Proof. By the construction of feature space, we can rewrite X and Y as follows:

$$X = \sum_{i=1}^{t-1}\mathbf{u}_i\mathbf{e}_i^T + \sum_{i=t}^{d}\tilde{\mathbf{u}}_i\mathbf{e}_i^T \qquad Y = \sum_{i=1}^{t-1}\mathbf{v}_i\mathbf{e}_i^T + \sum_{i=t}^{d}\tilde{\mathbf{v}}_i\mathbf{e}_i^T, \tag{11}$$

where for each $\tilde{\mathbf{u}}_i$, $\tilde{\mathbf{u}}_i^T\mathbf{u}_j = 0, \forall j$. Therefore, the trace norm of residual can be bounded by:

$$\begin{aligned}\|R - X\hat{M}Y^T\|_* &= \|\tilde{U}\tilde{U}^T R + R\tilde{V}\tilde{V}^T - \tilde{U}\tilde{U}^T R\tilde{V}\tilde{V}^T\|_*\\ &\leq 2\|\tilde{U}\tilde{U}^T U\Sigma V^T\|_* + \|U\Sigma V^T\tilde{V}\tilde{V}^T\|_*\\ &\leq 3\sum_{i=t}^{k}\sigma_i,\end{aligned}$$

where $\tilde{U}$, $\tilde{V}$ are the second term of X and Y in (11). Moreover, we have $\sigma_i = o(\sqrt{n})$ for all $i \geq t$. To see this, suppose $\sigma_p = \Omega(\sqrt{n})$ for any $t \leq p \leq k$, then:

$$\lim_{n\to\infty} \frac{\sigma_t}{\sqrt{n}} \geq \lim_{n\to\infty} \frac{\sigma_p}{\sqrt{n}} > 0,$$

leading a contradiction to the definition of σ_t. Therefore we can conclude:

$$\mathcal{N} = \|R - X\hat{M}Y^T\|_* \leq 3\sum_{i=t}^{k} \sigma_i \leq 3k \times o(\sqrt{n}) = o(\sqrt{n}),$$

and the Theorem is thus proved by plugging the above bound to Corollary 1. □

Proof of Theorem 3

Proof. We prove the Theorem by showing that the trace norm of $R - X\hat{M}Y^T$ will be $O((g(n) + h(n))\log n)$ in this scenario given that other dimensions (d and k) do not grow as a function of n. First, note that in this scenario, we can denote $X = U + \Delta U$ and $Y = V + \Delta V$, where $U \subseteq \text{col}(R), V \subseteq \text{row}(R)$ and $\Delta U, \Delta V$ are $g(n), h(n)$ column sparse respectively. The following Lemma then bounds the trace norm of $R - X\hat{M}Y^T$ in terms of ΔU and ΔV.

Lemma 6. *Let $\Delta U, \Delta V$ be defined as above. Then with high probability,*

$$\|R - X\hat{M}Y^T\|_* \leq c_1\xi_1\sqrt{\frac{k}{g(n)}}\|\Delta U^T R\|_* + c_2\xi_2\sqrt{\frac{k}{h(n)}}\|R\Delta V\|_* \tag{12}$$

with some universal constants c_1 and c_2.

Proof. Let $\Delta U = U_1\Sigma_1 V_1^T$ and $\Delta V = U_2\Sigma_2 V_2^T$ be the reduced SVD of the perturbation matrix $\Delta U, \Delta V$ accordingly. Then we have:

$$\begin{aligned}\|R - X\hat{M}Y^T\|_* &\leq \|U_1U_1^T R + RU_2U_2^T - U_1U_1^T RU_2U_2^T\|_* \\ &\leq 2\|U_1U_1^T R\|_* + \|RU_2U_2^T\|_* \\ &= 2\|\Delta U(V_1\Sigma_1^{-2}V_1^T)\Delta U^T R\|_* + \|R\Delta V(V_2\Sigma_2^{-2}V_2^T)\Delta V^T\|_*. \end{aligned}\tag{13}$$

For the first term of (13), using Hölder's inequality, we can upper bound it by:

$$\|\Delta U\|_2\|V_1\Sigma_1^{-2}V_1^T\|_2\|\Delta U^T R\|_* = \|\Delta U\|_2\|\Sigma_1^{-2}\|_2\|\Delta U^T R\|_*, \tag{14}$$

which suggests that we need to bound the largest and smallest singular values of ΔU to bound (14). Consider $\Delta U' \in \mathbb{R}^{g(n)\times k}$ to be the truncated ΔU where only non-zero rows in ΔU are left. The spectrum of $\Delta U'$ is same as ΔU. Moreover, its two norm can be bounded by:

$$\|\Delta U'\|_2 \leq \|\xi_1 E_1\|_2 \leq \xi_1\sqrt{kg(n)},$$

where $E_1 \in \mathbb{R}^{g(n)\times k}$ is the matrix with all entries are one. Also, using the result of [31], we can guarantee that with high probability $\sigma_k(\Delta U') \geq \Omega(\sqrt{g(n)} - \sqrt{k})$, which suggests w.h.p.:

$$\|\Sigma_1^{-2}\| = \frac{1}{\sigma_k(\Delta U)^2} = \frac{1}{\sigma_k(\Delta U')^2} \leq O(\frac{1}{g(n)}).$$

Thus, combining the above two bounds, the first term of (13) can be upper bounded by:

$$c_1\xi_1\sqrt{\frac{k}{g(n)}}\|\Delta U^T R\|_*,$$

with some universal constant c_1. Similarly, the second term of (13) can be upper bounded by $c_2\xi_2\sqrt{k/h(n)}\|R\Delta V\|_*$. The lemma is thus proved. □

Therefore, given Lemma 6, we now need to bound $\|\Delta U^T R\|_*$ and $\|R\Delta V\|_*$. We first focus on bounding the term $\|\Delta U^T R\|_*$. By $R = UV^T$ and the construction of U, V, we have:

$$\|\Delta U^T R\|_* = \|\Delta U^T UV^T\|_* \leq \|GV^T\|_*$$

where $G \in \mathbb{R}^{k\times k}$ with each entry in $G_{ij} \sim \xi_1 g(n)\mathcal{N}(0, \sigma^2)$. Thus, let $Z = GV^T, Z \in \mathbb{R}^{k\times n}$, then each entry $Z_{ij} \sim \xi_1 g(n)\frac{\sigma^2}{2}\chi_k^2$, where χ_k^2 is a chi-square distribution with degree of freedom k.

We next show that the trace norm of Z will be bounded in small enough order with high probability. To begin with, the following Lemma is used as an exponentially decreasing bound on the tail distribution of chi-square statistics.

Lemma 7 (Exponential Tail Bound of χ_k^2). *Let X be a random variable which follows χ_k^2. Then for any $t > 1$, we have:*

$$\Pr(X \geq tk) \leq \exp\left\{\frac{-k(\sqrt{(t-1)^2+1}-1)}{2}\right\}$$

This Lemma is a corollary of Lemma 1 in [23]. Given this lemma, we can now derive the following lemma to upper bound $\|\Delta U^T R\|_*$:

Lemma 8. *Let $\Delta U^T R \in \mathbb{R}^{k\times n}$ where ΔU and R are set as in Theorem 3. Then its trace norm can be upper bounded by:*

$$\|\Delta U^T R\|_* \leq C_1 k^{\frac{3}{2}} g(n)\sqrt{n}\log n$$

with probability at least $1 - kn^{-\frac{k-2}{2}}$.

Proof. Since $\|\Delta U^T R\|_* \leq \|Z\|_*$ where $Z_{ij} \sim \xi_1 g(n)\frac{\sigma^2}{2}\chi_k^2$, by applying Lemma 7 with $t = \log n$, we can guarantee that with probability at least $1 - n^{\frac{-k}{2}}$:

$$Z_{ij} \leq \xi_1 g(n)\frac{\sigma^2}{2}k\log n.$$

Thus, by applying union bound on each Z_{ij}, with probability at least $1 - kn^{-\frac{k-2}{2}}$:

$$\|\Delta U^T R\|_* \leq \|Z\|_* \leq \xi_1 g(n)\frac{\sigma^2}{2}k\log n\|E\|_*,$$

where $E \in \mathbb{R}^{k\times n}$ is a rank-1 matrix with all entries are 1. We can thus conclude the Lemma by the fact that $\|E\|_* = \|E\|_2 = \sqrt{nk}$. □

Similarly, by using the same proof steps, it could also be shown that $\|R\Delta V\|_* \leq C_2 k^{3/2} h(n)\sqrt{n}\log n$. Therefore, substituting above bounds back to Lemma 6, we obtain:

$$\begin{aligned}
\mathcal{N} &= \|R - X\hat{M}Y^T\|_* \\
&\leq c_1\xi_1\sqrt{\frac{k}{g(n)}}\|\Delta U^T R\|_* + c_2\xi_2\sqrt{\frac{k}{h(n)}}\|R\Delta V\|_* \\
&= O\big(\max(\sqrt{g(n)}, \sqrt{h(n)})\sqrt{n}\log n\big),
\end{aligned}$$

and the proof is thus completed by plugging this result into Corollary 1. □

Appendix C: More Synthetic experiments for DirtyIMC

Experiment on random orthogonal model

Here we conduct an experiment based on random orthogonal model stated in Theorem 2. We create a low rank matrix $R = U\Sigma V^T$ where $U, V \in \mathbb{R}^{n\times 20}$ are both random orthogonal matrix, and the singular values to be $\sqcup_{\alpha=1}^{10}\{\alpha n, \alpha\log n\}$, so there are 10 singular values have smaller growth rate $O(\log n)$. Follow Theorem 2, we construct X, Y by replacing the bottom 10 singular vectors in

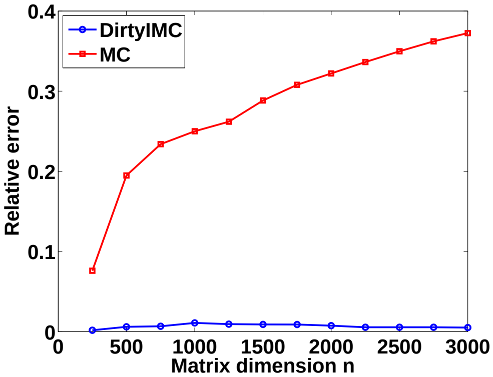


Figure 3: A synthetic experiment where noise only corrupts the insignificant part of true latent features (i.e. space spanned by smaller singular values). We see that in this case, given $O(n)$ observations, DirtyIMC could still recover the underlying matrix using sufficiently informative features, while matrix completion fails to recover as the error becomes unbounded with larger n. The result supports guarantee provided in Theorem 2.

U and V with bases orthogonal to U and V. We increase n from 250 to 3000, and for each n we randomly sample $m = 100n$ observations, apply our model and matrix completion to complete the matrix, and evaluate the recovered matrix using relative error. From Theorem 2, our DirtyIMC model should be able to approximately recover the matrix given $100n > o(n)$ observations, which is indeed true as Figure 3 suggests. As a comparison, standard matrix completion fails to recover the matrix with only $O(n)$ observations as n increases. This result empirically supports our theoretical analysis on the usefulness of noisy features.

Finer results for synthetic experiments in Section 5

Figure 4 and 5 show finer plots under each sparsity of observation ρ_s and feature noise level ρ_f.

Comparisons between DirtyIMC and Boosted IMC

As we mentioned in Appendix A, a recently proposed method "Boosted IMC" [33] could be viewed as a special case of our model, where their method is basically Algorithm 1 with $t_{max} = 1$, and in each subproblem they replace the trace norm regularized objective with matrix factorization objective. Here we compare our DirtyIMC (Algorithm 1) with Boosted IMC on synthetic datasets generated as same as Section 5 stated. We follow their implementation with rank of U, V, W, H are all set to be 40. The result is shown in Figure 6.

We observe that though Boosted IMC has a similar trend to DirtyIMC, in general, DirtyIMC performs better than Boosted IMC. However, Boosted IMC may be still good enough as an approximation of DirtyIMC in certain cases where efficiency is critical, since it only requires one iteration update of M and N.

Appendix D: Details for applying DirtyIMC to semi-supervised clustering

Here we follow the discussion in Section 5 for semi-supervised clustering. Suppose we are given m pairwise constraints describing similarity (or dissimilarity) of some pairs of items, then we can construct the following pairwise similarity matrix S as:

$$S_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are similar,} \\ 0, & \text{if } i \text{ and } j \text{ are dissimilar.} \end{cases}$$

Obviously, S has many missing entries since only $m \ll n^2$ pairwise constraints are known. In addition, ideally S should be a subset of observations sampled from UU^T, where $U \in \mathbb{R}^{n \times k}$ with each i-th column of U is an indicator vector of the i-th cluster. Therefore, one can try to recover

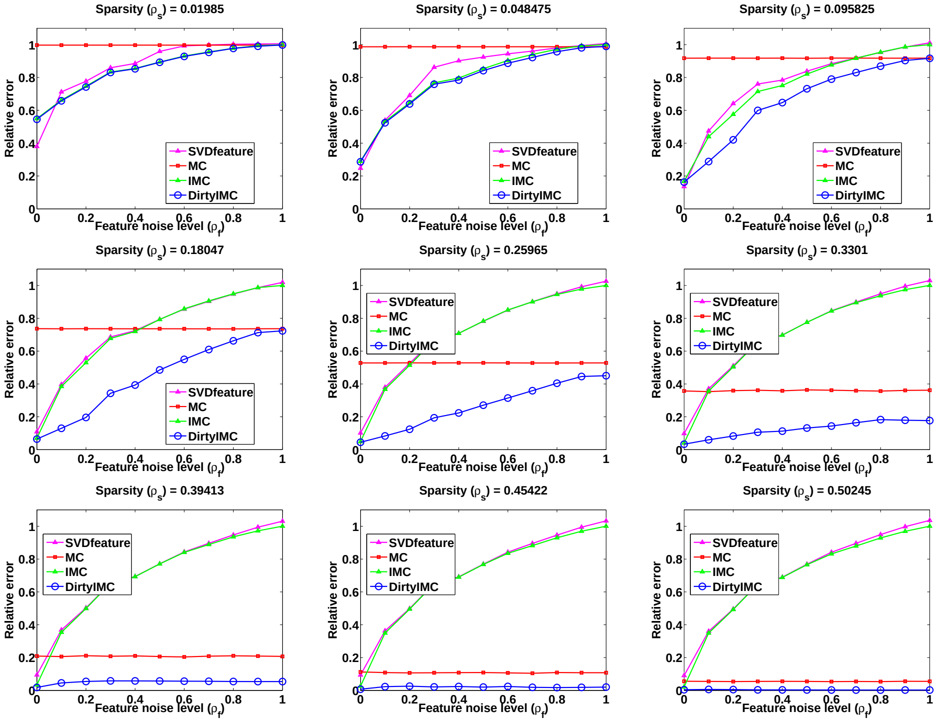


Figure 4: Finer results for synthetic experiments where completion methods are applied under different feature quality with a fixed ρ_s

(or complete) the matrix back with DirtyIMC objective, and the column space of recovered matrix, spanned by its top-k eigenvectors, will (ideally) reveal the indicator vectors. Our detailed algorithm is summarized in Algorithm 2.

One subtle yet critical issue in Algorithm 2 is to compute the top-k eigenvectors of recovered S (denoted as R). Note that after solving DirtyIMC objective, we are only given the low rank expression of N^* and M^*. Compute R explicitly and then compute its leading eigenvectors is expensive and not scalable. Therefore, we instead run subspace iteration on $N^* + ZM^*Z^T$ to solve for top-k eigenvectors efficiently. Also, since the resulting top-k eigenvectors are used for running k-means, we do not need to obtain a very accurate eigenvectors in this case. Therefore, parameters associated with precision (t_{max} and ϵ) could be set relatively loose for efficiency in practice.

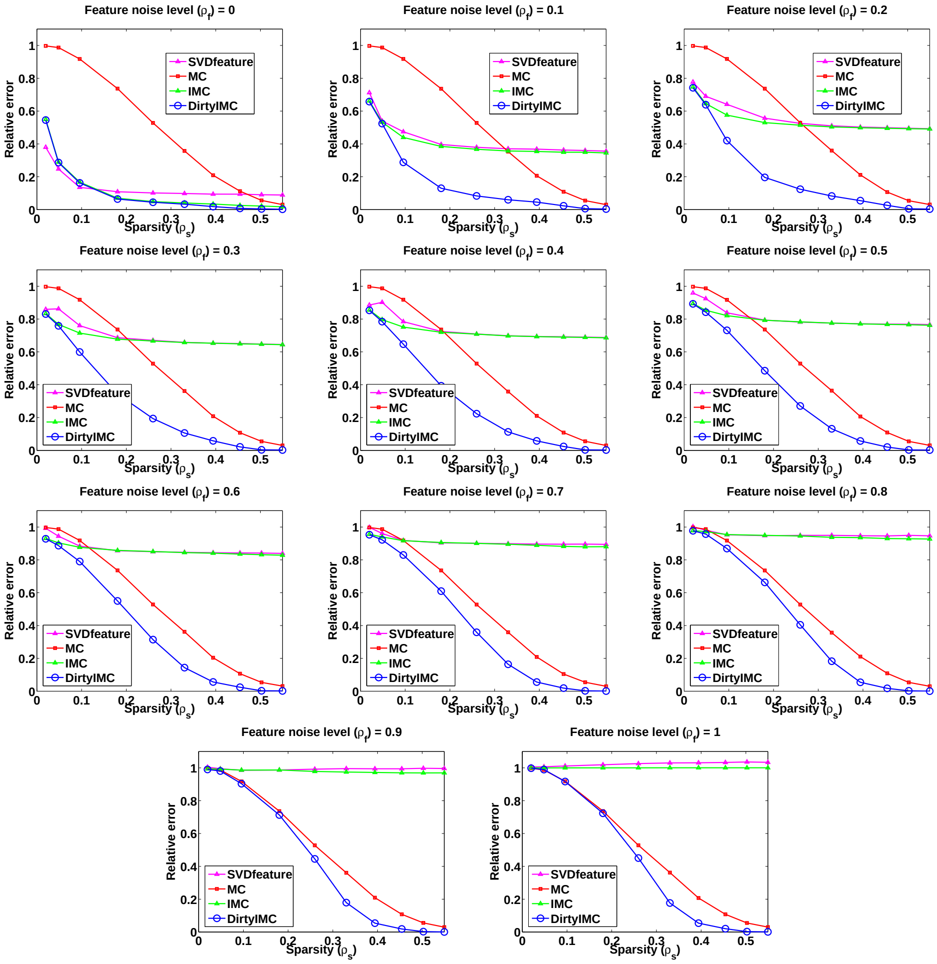
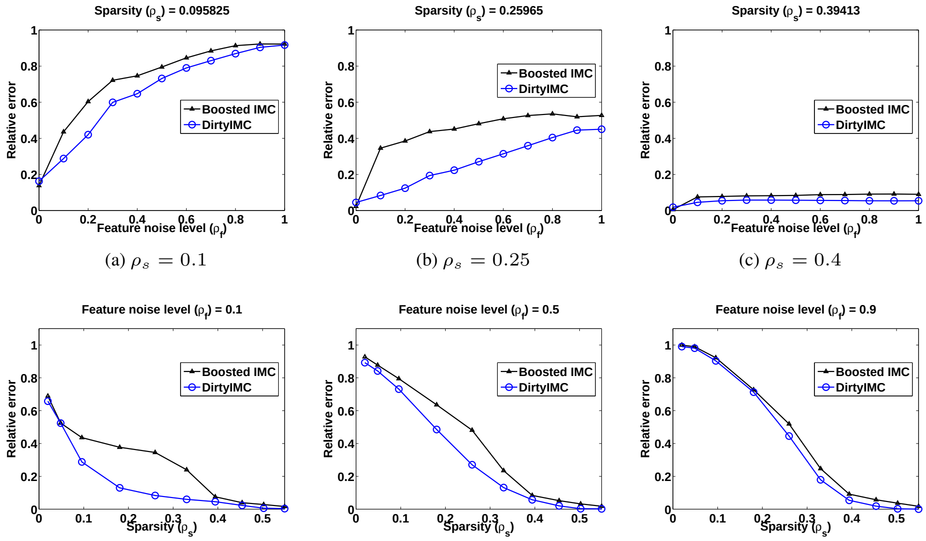


Figure 5: Finer results for synthetic experiments where completion methods are applied under different sparsity of observations with a fixed ρ_f



(a) $\rho_s = 0.1$

(b) $\rho_s = 0.25$

(c) $\rho_s = 0.4$

(d) $\rho_f = 0.1$

(e) $\rho_f = 0.5$

(f) $\rho_f = 0.9$

Figure 6: Performance of DirtyIMC and Boosted IMC (an approximation of DirtyIMC model) on synthetic datasets.

Algorithm 2 Semi-supervised clustering with DirtyIMC

```
**Input:** feature matrix $Z$, pairwise similarity matrix $S$, number of clusters $k$, regularization parameters $(\lambda_M, \lambda_N)$ in (2).
// Solve DirtyIMC objective with Algorithm 1.
$(M^*, N^*) \leftarrow \arg\min_{M,N} \sum_{(i,j)\in S}((ZMZ^T + N)_{ij} - S_{ij})^2 + \lambda_M \|M\|_* + \lambda_N \|N\|_*$
// Subspace iterations for finding top-$k$ eigenvectors.
$\epsilon \leftarrow 10^{-3}, t_{max} \leftarrow 10, t \leftarrow 1$
$[U_M, \Sigma_M, V_M] \leftarrow \text{SVD}(M^*)$
**initialize** $U_{(t)} \leftarrow \text{QR}(ZU_M, k)$
**while** $t \le t_{max}$ **do**
    $U_{(t+1)} \leftarrow \text{QR}(ZM^*Z^TU_{(t)} + N^*U_{(t)}, k)$
    $t \leftarrow t + 1$
    **if** $\sigma_k(U_{(t)}^T U_{(t+1)}) < \epsilon$ **then**
        **break**
    **end if**
**end while**
idx $\leftarrow$ kmeans$(U_{(t)}, k)$
**return** idx
```