Predictive Low Rank Matrix Learning under Partial Observations: Mixed-Projection ADMM

Dimitris Bertsimas DBERTSIM@MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Nicholas A. G. Johnson NAGJ@MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: TBD

Abstract

We study the problem of learning a partially observed matrix under the low rank assumption in the presence of fully observed side information that depends linearly on the true underlying matrix. This problem consists of an important generalization of the Matrix Completion problem, a central problem in Statistics, Operations Research and Machine Learning, that arises in applications such as recommendation systems, signal processing, system identification and image denoising. We formalize this problem as an optimization problem with an objective that balances the strength of the fit of the reconstruction to the observed entries with the ability of the reconstruction to be predictive of the side information. We derive a mixed-projection reformulation of the resulting optimization problem and present a strong semidefinite cone relaxation. We design an efficient, scalable alternating direction method of multipliers algorithm that produces high quality feasible solutions to the problem of interest. Our numerical results demonstrate that in the small rank regime ($k \leq 15$), our algorithm outputs solutions that achieve on average 79% lower objective value and 90.1% lower ℓ_2 reconstruction error than the solutions returned by the experiment-wise best performing benchmark method. The runtime of our algorithm is competitive with and often superior to that of the benchmark methods. Our algorithm is able to solve problems with $n = 10000$ rows and $m = 10000$ columns in less than a minute.

Keywords: Matrix Completion; Rank; Mixed-Projection; ADMM;

1. Introduction

In many real world applications, we are faced with the problem of recovering a (often large) matrix from a (often small) subset of its entries. This problem, known as Matrix Completion (MC), has gained significant attention due to its broad range of applications in areas such as signal processing (Candes and Plan, 2010), system identification (Liu, 2019) and image denoising (Ji et al., 2010). The fundamental task in MC is to accurately reconstruct the missing entries of a matrix given a limited number of observed entries. The challenge is particularly pronounced when the number of observed entries is small relatively to the dimension of the matrix, yet this is the common scenario in practice.

One of the most prominent uses of MC is in recommendation systems, where the goal is to predict user preferences for items (e.g., movies, products) based on a partially observed user-

item rating matrix (Ramlatchan et al., 2018). The Netflix Prize competition highlighted the potential of MC techniques, where the objective was to predict missing ratings in a user-movie matrix to improve recommendation accuracy (Bell and Koren, 2007). The success of such systems hinges on the assumption that the underlying rating matrix is low rank, meaning that the preferences of users can be well-approximated by a small number of factors. Indeed, it has been well studied that many real world datasets are low rank (Udell and Townsend, 2019).

In many practical applications, in addition to a collection of observed matrix entries we additionally have access to auxiliary side information that can be leveraged when performing the reconstruction. For example, in a recommendation system, side information might consist of social network data or item attributes. The vast majority of existing approaches to MC in the presence of side information incorporate the side information by making additional structural restrictions on the reconstructed matrix beyond the usual low rank assumption (see, for example, Xu et al. (2013); Chiang et al. (2015); Bertsimas and Li (2020)). In this work, we take an alternate approach by assuming that the side information can be well modelled as a linear function of the underlying full matrix. In this setting, the side information can be thought of as labels for a regression problem where the unobserved matrix consists of the regression features. This assumption is in keeping with ideas from the predictive low rank kernel learning literature (Bach and Jordan, 2005) (note however that low rank kernel learning assumes a fully observed input matrix).

Formally, let $\Omega \subseteq [n] \times [m]$ denote a collection of revealed entries of a partially observed matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, let $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ denote a matrix of side information and let k denote a specified target rank. We consider the problem given by

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{R}^{n \times m}, \boldsymbol{\alpha} \in \mathbb{R}^{m \times d}} \quad & \sum_{(i,j) \in \Omega} (X_{ij} - A_{ij})^2 + \lambda \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \gamma \|\boldsymbol{X}\|_\star \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{X}) \leq k, \end{aligned} \tag{1}$$

where $\lambda, \gamma > 0$ are hyperparameters that in practice can either take a default value or can be cross-validated by minimizing a validation metric (Owen and Perry, 2009) to obtain strong out-of-sample performance (Bousquet and Elisseeff, 2002). We assume that the ground truth matrix $\boldsymbol{A}$ has low rank and that the side information can be well approximated as $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{N}$ for some weighting matrix $\boldsymbol{\alpha}$ and noise matrix $\boldsymbol{N}$. The first term in the objective function of (1) measures how well the observed entries of the unknown matrix are fit by the estimated matrix $\boldsymbol{X}$, the second term of the objective function measures how well the side information $\boldsymbol{Y}$ can be represented as a linear function of the estimated matrix $\boldsymbol{X}$ and the final term of the objective is a regularization term. To the best of our knowledge, Problem (1) has not previously been directly studied despite its very natural motivation.

1.1 Contribution and Structure

In this paper, we tackle (1) by developing novel mixed-projection optimization techniques (Bertsimas et al., 2022). We show that solving (1) is equivalent to solving an appropriately defined robust optimization problem. We develop an exact reformulation of (1) by combining a parametrization of the $\boldsymbol{X}$ decision variable as the product of two low rank factors with the introduction of a projection matrix to model the column space of $\boldsymbol{X}$. We derive a

semidefinite cone convex relaxation for our mixed-projection reformulation and we present an efficient, scalable alternating direction method of multipliers (ADMM) algorithm that produces high quality feasible solutions to (1). Our numerical results show that across all experiments in the small rank regime ($k \leq 15$), our algorithm outputs solutions that achieve on average 79% lower objective value in (1) and 90.1% lower ℓ_2 reconstruction error than the solutions returned by the experiment-wise best performing benchmark method. For the 5 experiments with $k > 15$, the only benchmark that returns a solution with superior quality than that returned by our algorithm takes on average 3 times as long to execute. The runtime of our algorithm is competitive with and often superior to that of the benchmark methods. Our algorithm is able to solve problems with $n = 10000$ rows and $m = 10000$ columns in less than a minute.

The rest of the paper is laid out as follows. In Section 2, we review previous work that is closely related to (1). In Section 3, we study (1) under a robust optimization lens and investigate formulating (1) as a two stage optimization problem where the inner stage is a regression problem that can be solved in closed form. We formulate (1) as a mixed-projection optimization problem in Section 4 and present a natural convex relaxation. In Section 5, we present and rigorously study our ADMM algorithm. Finally, in Section 6 we investigate the performance of our algorithm against benchmark methods on synthetic data.

Notation: We let nonbold face characters such as b denote scalars, lowercase bold faced characters such as $\boldsymbol{x}$ denote vectors, uppercase bold faced characters such as $\boldsymbol{X}$ denote matrices, and calligraphic uppercase characters such as $\mathcal{Z}$ denote sets. We let $[n]$ denote the set of running indices $\{1, ..., n\}$. We let $\mathbf{0}_n$ denote an n-dimensional vector of all 0's, $\mathbf{0}_{n\times m}$ denote an $n \times m$-dimensional matrix of all 0's, and $\boldsymbol{I}_n$ denote the $n \times n$ identity matrix. We let $\mathcal{S}^n$ denote the cone of $n \times n$ symmetric matrices and $\mathcal{S}^n_+$ denote the cone of $n \times n$ positive semidefinite matrices

2. Literature Review

In this section, we review a handful of notable approaches from the literature that have been employed to solve MC and to solve general low rank optimization problems. As an exhaustive literature review of MC methods is outside of the scope of this paper, we focus our review on a handful of well studied approaches which we will employ as benchmark methods in this work. We additionally give an overview of the ADMM algorithmic framework which is of central relevance to this work. For a more detailed review of the MC literature, we refer the reader to Ramlatchan et al. (2018) and Nguyen et al. (2019).

2.1 Matrix Completion Methods

2.1.1 Iterative-SVD

Iterative-SVD is an expectation maximization style algorithm (Dempster et al., 1977) that generates a solution to the MC problem by iteratively computing a singular value decomposition (SVD) of the current iterate and estimating the missing values by performing a regression against the low rank factors returned by SVD (Troyanskaya et al., 2001). This is one of a handful of methods in the literature that leverage the SVD as their primary al-

gorithmic workhorse (Billsus et al., 1998; Sarwar et al., 2000). Concretely, given a partially observed matrix $\{X_{ij}\}_{(i,j)\in\Omega}$ where $\Omega \subseteq [n] \times [m]$ and a target rank $k \in \mathbf{N}_+$, Iterative-SVD proceeds as follows:

1. Initialize the iteration count $t \leftarrow 0$ and initialize missing entries of $X_{ij}, (i,j) \notin \Omega$ with the row average $X_{ij} = \frac{\sum_{l:(i,l)\in\Omega} X_{il}}{|\{l:(i,l)\in\Omega\}|}$.

2. Compute a rank k SVD $\boldsymbol{X}_t = \boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^T$ of the current iterate where $\boldsymbol{U}_t \in \mathbb{R}^{n\times k}, \boldsymbol{\Sigma}_t \in \mathbb{R}^{k\times k}, \boldsymbol{V}_t \in \mathbb{R}^{m\times k}$.

3. For each $(i,j) \notin \Omega$, estimate the missing value $(\boldsymbol{X}_{t+1})_{ij}$ by regressing all other entries in row i against all except the j^{th} row of $\boldsymbol{V}_t$. Concretely, letting $\tilde{\boldsymbol{x}} = (\boldsymbol{X}_t)_{i,\star\setminus j} \in \mathbb{R}^{m-1}$ denote the column vector consisting of the i^{th} row of $\boldsymbol{X}_t$ excluding the j^{th} entry, letting $\tilde{\boldsymbol{V}} = (\boldsymbol{V}_t)_{\star\setminus j,\star} \in \mathbb{R}^{(m-1)\times k}$ denote the matrix formed by eliminating the j^{th} row from $\boldsymbol{V}_t$ and letting $\hat{\boldsymbol{v}} = (\boldsymbol{V}_t)_{j,\star} \in \mathbb{R}^k$ denote the column vector consisting of the j^{th} row of $\boldsymbol{V}_t$, we set $(\boldsymbol{X}_{t+1})_{ij} = \hat{\boldsymbol{v}}^T(\tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{V}})^{-1}\tilde{\boldsymbol{V}}^T\tilde{\boldsymbol{x}}$.

4. Terminate if the total change between $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t+1}$ is less than 0.01. Otherwise, increment t and return to Step 2.

2.1.2 Soft-Impute

Soft-Impute is a convex relaxation inspired algorithm that leverages the nuclear norm as a low rank inducing regularizer (Mazumder et al., 2010). This approach is one of a broad class of methods that tackle MC from a nuclear norm minimization lens (Fazel, 2002; Candès and Tao, 2010; Candes and Recht, 2012). Seeking a reconstruction with minimum nuclear norm is typically motivated by the observation that the nuclear norm ball given by $\mathcal{B} = \{\boldsymbol{X} \in \mathbb{R}^{n\times n} : \|\boldsymbol{X}\|_\star \leq k\}$ is the convex hull of the nonconvex set $\mathcal{X} = \{\boldsymbol{X} \in \mathbb{R}^{n\times n} : \text{rank}(\boldsymbol{X}) \leq k, \|\boldsymbol{X}\|_\sigma \leq 1\}$, where $\|\cdot\|_\sigma$ denotes the spectral norm. Moreover, several conditions have been established under which nuclear norm minimization methods are guaranteed to return the ground truth matrix (Candès and Tao, 2010; Candes and Recht, 2012) though such conditions tend to be strong and hard to verify in practice. Soft-Impute proceeds by iteratively replacing the missing elements of the matrix with those obtained from a soft thresholded low rank singular value decomposition. Accordingly, similarly to Iterative-SVD, Soft-Impute relies on the computation of a low rank SVD as the primary algorithmic workhorse. The approach relies on the result that for an arbitrary matrix $\boldsymbol{X}$, the solution of the problem $\min_{\boldsymbol{Z}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Z}\|_F^2 + \lambda\|\boldsymbol{Z}\|_\star$ is given by $\hat{\boldsymbol{Z}} = S_\lambda(\boldsymbol{X})$ where $S_\lambda(\cdot)$ denotes the soft-thresholding operation (Donoho et al., 1995). Explicitly, Soft-Impute proceeds as follows for a given regularization parameter $\lambda > 0$ and termination threshold $\epsilon > 0$:

1. Initialize the iteration count $t \leftarrow 0$ and initialize $\boldsymbol{Z}_t = \boldsymbol{0}_{n\times m}$.

2. Compute $\boldsymbol{Z}_{t+1} = S_\lambda(P_\Omega(\boldsymbol{X}) + P_\Omega^\perp(\boldsymbol{Z}_t))$ where $P_\Omega(\cdot)$ denotes the operation that projects onto the revealed entries of $\boldsymbol{X}$ while $P_\Omega^\perp(\cdot)$ denotes the operation that projects onto the missing entries of $\boldsymbol{X}$.

3. Terminate if $\frac{\|\boldsymbol{Z}_t-\boldsymbol{Z}_{t+1}\|_F^2}{\|\boldsymbol{Z}_t\|_F^2}$. Otherwise, increment t and return to Step 2.

2.1.3 Fast-Impute

Fast-Impute is a projected gradient descent approach to MC that has desirable global convergence properties (Bertsimas and Li, 2020). Fast-Impute belongs to the broad class of methods that solve MC by factorizing the target matrix as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n\times k}, \boldsymbol{V} \in \mathbb{R}^{m\times k}$ and performing some variant of gradient descent (or alternating minimization) on the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ (Koren et al., 2009; Jain and Netrapalli, 2015; Zheng and Lafferty, 2016; Jin et al., 2016). We note that we leverage this common factorization in the approach to (1) presented in this work. Gradient descent based methods have shown great success. Despite the non-convexity of the factorization, it has been shown that in many cases gradient descent and its variants will nevertheless converge to a globally optimal solution (Chen and Wainwright, 2015; Ge et al., 2015; Sun and Luo, 2016; Ma et al., 2018; Bertsimas and Li, 2020). Fast-Impute takes the approach of expressing $\boldsymbol{U}$ as a closed form function of $\boldsymbol{V}$ after performing the facorization and directly performs projected gradient descent updates on $\boldsymbol{V}$ with classic Nesterov acceleration (Nesterov, 1983). Moreover, to enhance scalability of their method, Bertsimas and Li (2020) design a stochastic gradient extension of Fast-Impute that estimates the gradient at each update step by only considering a sub sample of the rows and columns of the target matrix.

2.2 Low Rank Optimization Methods

2.2.1 ScaledGD

ScaledGD is a highly performant method to obtain strong solutions to low rank matrix estimation problems that take the following form:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times m}} f(\boldsymbol{X}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2 \quad \text{s.t. } \operatorname{rank}(\boldsymbol{X}) \leq k,$$

where $\mathcal{A}(\cdot) : \mathbb{R}^{n\times m} \to \mathbb{R}^l$ models some measurement process and we have $\boldsymbol{y} \in \mathbb{R}^l$ (Tong et al., 2021). ScaledGD proceeds by factorizing the target matrix as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ and iteratively performing gradient updates on the low rank factors $\boldsymbol{U}, \boldsymbol{V}$ after preconditioning the gradients with an adaptive matrix that is efficient to compute. Doing so yields a linear convergence rate that is notably independent of the condition number of the low rank matrix. In so doing, ScaledGD combines the desirable convergence rate of alternating minimization with the desirable low per-iteration cost of gradient descent. Explicitly, letting $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}) = f(\boldsymbol{U}\boldsymbol{V}^T)$, ScaledGD updates the low rank factors as:

$$\boldsymbol{U}_{t+1} \leftarrow \boldsymbol{U}_t - \eta\nabla_{\boldsymbol{U}}\mathcal{L}(\boldsymbol{U}_t, \boldsymbol{V}_t)(\boldsymbol{V}^T\boldsymbol{V})^{-1},$$

$$\boldsymbol{V}_{t+1} \leftarrow \boldsymbol{V}_t - \eta\nabla_{\boldsymbol{V}}\mathcal{L}(\boldsymbol{U}_t, \boldsymbol{V}_t)(\boldsymbol{U}^T\boldsymbol{U})^{-1},$$

where $\eta > 0$ denotes the step size.

2.2.2 Mixed-Projection Conic Optimization

Mixed-projection conic optimization is a recently proposed modelling and algorithmic framework designed to tackle a broad class of matrix optimization problems (Bertsimas et al., 2022, 2023c). Specifically, this approach considers problems that have the following form:

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{R}^{n\times m}} \quad & \langle \boldsymbol{C}, \boldsymbol{X}\rangle + \lambda \cdot \text{rank}(\boldsymbol{X}) + \Omega(\boldsymbol{X}) \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{X} = \boldsymbol{B},\ \text{rank}(\boldsymbol{X}) \leq k,\ \boldsymbol{X} \in \mathcal{K}, \end{aligned} \tag{2}$$

where $\boldsymbol{C} \in \mathbb{R}^{n\times m}$ is a cost matrix, $\lambda > 0$, $k \in \mathbb{N}_+$, $\boldsymbol{A} \in \mathbb{R}^{l\times n}, \boldsymbol{B} \in \mathbb{R}^{l\times m}, \mathcal{K}$ denotes a proper cone in the sense of Boyd et al. (2004) and $\Omega(\cdot)$ is a frobenius norm regularization function or a spectral norm regularization function of the input matrix. The main workhorse of mixed-projection conic optimization is the use of a projection matrix to cleverly model the rank terms in (2). This can be viewed as the matrix generalization of using binary variables to model the sparsity of a vector in mixed-integer optimization. Bertsimas et al. (2022) show that for an arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, we have

$$\text{rank}(\boldsymbol{X}) \leq k \iff \exists \boldsymbol{P} \in \mathbb{S}^n : \boldsymbol{P}^2 = \boldsymbol{P}, \text{Tr}(\boldsymbol{P}) \leq k, \boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}.$$

Introducing projection matrices allows the rank functions to be eliminated from (2) at the expense of introducing non convex quadratic equality constraints. From here, most existing works that leverage mixed-projection conic optimization have either focused on obtaining strong semidefinite based convex relaxations (Bertsimas et al., 2022, 2023c) or have focused on obtaining certifiably optimal solutions for small and moderately sized problem instances (Bertsimas et al., 2023a,b). In this work, we leverage the mixed-projection framework to scalably obtain high quality solutions to large problem instances.

2.3 Alternating Direction Method of Multipliers

Alternating direction method of multipliers (ADMM) is an algorithm that was originally designed to solve linearly constrained convex optimization problems of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^m} \quad f(\boldsymbol{x}) + g(\boldsymbol{z}) \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c}, \tag{3}$$

where we have $\boldsymbol{A} \in \mathbb{R}^{l\times n}, \boldsymbol{B} \in \mathbb{R}^{l\times m}, \boldsymbol{c} \in \mathbb{R}^l$ and the functions f and g are assumed to be convex (Boyd et al., 2011). The main benefit of ADMM is that it can combine the decomposition benefits of dual ascent with the desirable convergence properties of the method of multipliers. Letting $\boldsymbol{y} \in \mathbb{R}^l$ denote the dual variable, the augmented lagrangian of (3) is given by

$$\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \boldsymbol{y}^T(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}) + \frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}\|_2^2,$$

where $\rho > 0$ is the augmented lagrangian parameter. ADMM then proceeds by iteratively updating the primal variable $\boldsymbol{x}$, updating the primal variable $\boldsymbol{z}$ and taking a gradient ascent step on the dual variable $\boldsymbol{y}$. Explicitly, ADMM consists of the following updates:

1. $\boldsymbol{x}_{t+1} \leftarrow \text{argmin}_{\boldsymbol{x}}\, \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}_t, \boldsymbol{y}_t)$,
2. $\boldsymbol{z}_{t+1} \leftarrow \text{argmin}_{\boldsymbol{z}}\, \mathcal{L}(\boldsymbol{x}_{t+1}, \boldsymbol{z}, \boldsymbol{y}_t)$,
3. $\boldsymbol{y}_{t+1} \leftarrow \boldsymbol{y}_t + \rho(\boldsymbol{A}\boldsymbol{x}_{t+1} + \boldsymbol{B}\boldsymbol{z}_{t+1} - \boldsymbol{c})$.

Under very mild regularity conditions on f, g and $\mathcal{L}$, it is well known that ADMM is guaranteed to produce a sequence of primal iterates that converges to the optimal value of (3) and a sequence of dual iterates that converge to the optimal dual variable (note that there is no guarantee of primal variable convergence) (Boyd et al., 2011). Importantly, although ADMM was originally designed for linearly constrained convex optimization, it has often been applied to non convex optimization problems and yielded empirically strong results (Xu et al., 2016). This observation has motivated work to explore the theoretical convergence behavior of ADMM and its variants on specific classes of non convex optimization problems (Guo et al., 2017; Wang et al., 2019; Wang and Zhao, 2021).

3. Formulation Properties

In this section, we rigorously investigate certain key features of (1). Specifically, we establish an equivalence between (1) and an appropriately defined robust optimization problem. Moreover, we illustrate that (1) can be reduced to an optimization problem over only $\boldsymbol{X}$ and establish that the resulting objective function is not convex, not concave and non-smooth. Finally, we study how efficient evaluations of the reduced problem objective function can be performed.

3.1 Equivalence Between Regularization and Robustness

Real-world datasets frequently contain inaccuracies and missing values, which hinder the ability of machine learning models to generalize effectively to new data when these inconsistencies are not appropriately modelled. Consequently, robustness is a crucial quality for machine learning models, both in theory and application (Xu et al., 2009; Bertsimas and den Hertog, 2020). In this section, we show that our regularized problem (1) can be viewed as a robust optimization (RO) problem. This finding justifies the inclusion of the nuclear norm regularization term in (1) and is in a similar flavor as known results from the robust optimization literature in the case of vector (Bertsimas and Copenhaver, 2018) and matrix (Bertsimas et al., 2023a) problems.

Proposition 1 *Problem* (1) *is equivalent to the following robust optimization problem:*

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{R}^{n\times m},\boldsymbol{\alpha}\in\mathbb{R}^{m\times d}} \max_{\boldsymbol{\Delta}\in\mathcal{U}} \quad & \sum_{(i,j)\in\Omega} (X_{ij}-A_{ij})^2 + \lambda\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\alpha}\|_F^2 + \langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle \\ s.t. \quad & rank(\boldsymbol{X}) \leq k_0, \end{aligned} \tag{4}$$

where $\mathcal{U} = \{\boldsymbol{\Delta} \in \mathbb{R}^{n\times m} : \|\boldsymbol{\Delta}\|_\sigma \leq \gamma\}$

Proof To establish this result, it suffices to argue that $\max_{\boldsymbol{\Delta}\in\mathcal{U}}\langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle = \gamma\|\boldsymbol{X}\|_\star$. This equivalence follows immediately from the fact that the nuclear norm is dual to the spectral norm. So as to keep this manuscript self contained, we present a proof of this equivalence below.

Consider any matrix $\bar{\boldsymbol{\Delta}} \in \mathbb{R}^{n\times m}$ such that $\|\bar{\boldsymbol{\Delta}}\|_\sigma \leq \gamma$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a singular value decomposition of $\boldsymbol{X}$ where we let $r = \text{rank}(\boldsymbol{X})$ and we have $\boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}, \boldsymbol{V} \in$

$\mathbb{R}^{m\times r}$. We have

$$\langle \boldsymbol{X}, \bar{\boldsymbol{\Delta}}\rangle = \mathrm{Tr}(\bar{\boldsymbol{\Delta}}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T) = \mathrm{Tr}(\boldsymbol{V}^T\bar{\boldsymbol{\Delta}}^T\boldsymbol{U}\boldsymbol{\Sigma}) = \langle \boldsymbol{U}^T\bar{\boldsymbol{\Delta}}\boldsymbol{V}, \boldsymbol{\Sigma}\rangle = \sum_{i=1}^r \Sigma_{ii}(\boldsymbol{U}^T\bar{\boldsymbol{\Delta}}\boldsymbol{V})_{ii}$$
$$= \sum_{i=1}^r \Sigma_{ii}U_i^T\bar{\boldsymbol{\Delta}}V_i \leq \sum_{i=1}^r \Sigma_{ii}\sigma_1(\bar{\boldsymbol{\Delta}}) \leq \gamma\sum_{i=1}^r \Sigma_{ii} = \gamma\|\boldsymbol{X}\|_\star$$

where we have used the fact that $\boldsymbol{\Sigma}$ is a diagonal matrix and the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ have unit length. Thus, we have shown that $\gamma\|\boldsymbol{X}\|_\star$ is an upper bound for $\max_{\boldsymbol{\Delta}\in\mathcal{U}}\langle \boldsymbol{X}, \boldsymbol{\Delta}\rangle$. To show that the upper bound is always achieved, consider the matrix $\tilde{\boldsymbol{\Delta}} = \gamma\boldsymbol{U}\boldsymbol{V}^T \in \mathbb{R}^{n\times m}$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are taken from a singular value decomposition of $\boldsymbol{X}$. Observe that

$$\|\tilde{\boldsymbol{\Delta}}\|_\sigma = \gamma\|\boldsymbol{U}\boldsymbol{V}^T\|_\sigma \leq \gamma \implies \tilde{\boldsymbol{\Delta}} \in \mathcal{U}.$$

We conclude by noting that

$$\langle \boldsymbol{X}, \tilde{\boldsymbol{\Delta}}\rangle = Tr(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\gamma\boldsymbol{U}\boldsymbol{V}^T) = \gamma Tr(\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}) = \gamma\mathrm{Tr}(\boldsymbol{I}\boldsymbol{\Sigma}\boldsymbol{I}) = \gamma\|\boldsymbol{X}\|_\star.$$

■

Proposition 1 implies that solving the nuclear norm regularized (1) is equivalent to solving an unregularized robust optimization problem that protects against adversarial perturbations that are bounded in spectral norm. This result is not surprising given the duality of norms, yet is nevertheless insightful.

3.2 A Partial Minimization

Let $g(\boldsymbol{X}, \boldsymbol{\alpha})$ denote the objective function of (1). Note that $g(\boldsymbol{X}, \boldsymbol{\alpha})$ is bi-convex in $(\boldsymbol{X}, \boldsymbol{\alpha})$ but is not jointly convex due to the product $\boldsymbol{X}\boldsymbol{\alpha}$. Observe that we can simplify (1) by performing a partial minimization in $\boldsymbol{\alpha}$. For any $\boldsymbol{X}$, the problem in $\boldsymbol{\alpha}$ requires finding the unconstrained minimum of a convex quadratic function. The gradient of g with respect to $\boldsymbol{\alpha}$ is given by $\nabla_{\boldsymbol{\alpha}}g(\boldsymbol{X}, \boldsymbol{\alpha}) = 2\lambda\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{Y})$. Setting $\nabla_{\boldsymbol{\alpha}}g(\boldsymbol{X}, \boldsymbol{\alpha})$ to 0 yields $\boldsymbol{\alpha}^\star = (\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\boldsymbol{Y}$ as a minimizer of g over $\boldsymbol{\alpha}$. Letting $f(\boldsymbol{X})$ correspond to the partially minimized objective function of (1), we have

$$f(\boldsymbol{X}) = \min_{\boldsymbol{\alpha}} g(\boldsymbol{X}, \boldsymbol{\alpha}) = \sum_{(i,j)\in\Omega}(X_{ij} - A_{ij})^2 + \lambda\|(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T)\boldsymbol{Y}\|_F^2 + \gamma\|\boldsymbol{X}\|_\star$$
$$= \sum_{(i,j)\in\Omega}(X_{ij} - A_{ij})^2 + \lambda\mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T)\boldsymbol{Y}\big) + \gamma\|\boldsymbol{X}\|_\star$$

We note that $\boldsymbol{\alpha}^\star$ corresponds to the well studied ordinary least squares solution. When $\boldsymbol{X}^T\boldsymbol{X}$ has full rank, $\boldsymbol{\alpha}$ is the unique minimizer of g. If $\boldsymbol{X}^T\boldsymbol{X}$ is rank deficient, $\boldsymbol{\alpha}^\star$ corresponds to the minimizer with minimum norm.

Though we have simplified the objective function of (1), $f(\boldsymbol{X})$ is not a particularly well behaved function. We formalize this statement in Proposition (2).

Proposition 2 *The function $f(\boldsymbol{X})$ is in general neither convex nor concave and is non-smooth.*

Proof To illustrate that $f(\boldsymbol{X})$ is in general neither convex nor concave, suppose that $\Omega = \emptyset$, $n = 2$ and $m = d = \lambda = \gamma = 1$. In this setting, we have $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{2\times 1}$. Assuming that $\boldsymbol{x} \neq \boldsymbol{0}_2$, we can write the objective function as

$$\begin{aligned} f(\boldsymbol{x}) &= \mathrm{Tr}(\boldsymbol{y}^T(\boldsymbol{I}_2 - \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T)\boldsymbol{y}) + \|\boldsymbol{x}\|_\star \\ &= \boldsymbol{y}^T\boldsymbol{y} - \frac{\boldsymbol{y}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{x}} + \|\boldsymbol{x}\|_2 \\ &= \boldsymbol{y}^T\boldsymbol{y} - \frac{(\boldsymbol{y}^T\boldsymbol{x})^2}{\boldsymbol{x}^T\boldsymbol{x}} + \sqrt{\boldsymbol{x}^T\boldsymbol{x}}. \end{aligned}$$

For $\boldsymbol{x} = \boldsymbol{0}_2$, the objective value $f(\boldsymbol{0}_2)$ is equal to $\boldsymbol{y}^T\boldsymbol{y}$. Let $\boldsymbol{y} = \boldsymbol{1}_2$ and consider the line in $\boldsymbol{R}^2$ defined by $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^2 : x_2 = x_1 + 1\}$. The restriction of $f(\boldsymbol{x})$ to the line defined by $\mathcal{X}$ is a univariate function given by

$$f_{\mathcal{X}}(t) = 2 - \frac{(2t+1)^2}{2t^2+2t+1} + \sqrt{2t^2+2t+1}$$

where $t \in \mathbb{R}$ is a dummy variable. Observe that we have $f_{\mathcal{X}}(-1) = f_{\mathcal{X}}(0) = 2, f_{\mathcal{X}}(-0.5) = 2 + \frac{\sqrt{2}}{2}$ and $f_{\mathcal{X}}(-4) = f_{\mathcal{X}}(3) = 5.04$. Thus, the point $(-0.5, f_{\mathcal{X}}(0.5))$ lies above the chord connecting $(-1, f_{\mathcal{X}}(-1))$ and $(0, f_{\mathcal{X}}(0))$, so $f_{\mathcal{X}}(t)$ is not a convex function. Moreover, the point $(-1, f_{\mathcal{X}}(-1))$ lies below the chord connecting $(-4, f_{\mathcal{X}}(-4))$ and $(-0.5, f_{\mathcal{X}}(-0.5))$, so $f_{\mathcal{X}}(t)$ is not a concave function. Since a function is convex (respectively concave) if and only if its restriction to every line is convex (respectively concave), we have established that $f(\boldsymbol{X})$ is neither convex nor concave since $f_{\mathcal{X}}(t)$ is neither convex nor concave. To conclude the proof of Proposition 2, note that the non-smooth property of $f(\boldsymbol{X})$ follows immediately from the non-smooth property of the nuclear norm function. ■

Although the above closed form partial minimization in $\boldsymbol{\alpha}$ eliminates $m \times d$ variables form (1), this comes at the expense of introducing a $m \times m$ matrix pseudo-inverse term into the objective function which can be computationally expensive to evaluate. Efficient evaluation of an objective function is crucial in many optimization problems to quickly measure solution quality. A plethora of modern optimization techniques require iterative objective function evaluations. As a result, the computational cost of evaluating an objective function can quickly become the bottleneck of an algorithm's complexity. Directly evaluating $f(\boldsymbol{X})$ naively requires $O(|\Omega|)$ operations for the first term, $O(m^2n + m^3 + n^2d)$ for the second term (forming the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is $O(m^2n)$, taking the pseudo-inverse is $O(m^3)$, computing the products involving $\boldsymbol{Y}$ is $O(n^2d)$) and requires $O(mn\min(m,n))$ for the third term (the nuclear norm can be evaluated by computing a singular value decomposition of $\boldsymbol{X}$). We observe that computing the second term of $f(\boldsymbol{X})$ involving the pseudo-inverse dominates the complexity calculation. Indeed, the overall complexity of evaluating $f(\boldsymbol{X})$ naively is $O(m^2n + m^3 + n^2d)$.

Fortunately, it is possible to make evaluations of $f(\boldsymbol{X})$ without explicitly forming the product $\boldsymbol{X}^T\boldsymbol{X}$ or taking a pseudo-inverse. Proposition 3 illustrates that it suffices (in terms of computational complexity) to take a singular value decomposition of $\boldsymbol{X}$. Moreover, a large class of optimization algorithms require only function evaluations for feasible solutions. If we consider only those values of $\boldsymbol{X}$ that are feasible to (1), it is sufficient (in terms of

computational complexity) to take a rank k truncated singular value decomposition of $\boldsymbol{X}$ to make functions evaluations of $f(\boldsymbol{X})$.

Proposition 3 *The function $f(\boldsymbol{X})$ can equivalently be written as*

$$f(\boldsymbol{X}) = \sum_{(i,j)\in\Omega} (X_{ij} - A_{ij})^2 + \lambda Tr\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{Y}\big) + \gamma\|\boldsymbol{X}\|_\star,$$

where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is a singular value decomposition of $\boldsymbol{X}$ where we let $r = rank(\boldsymbol{X})$ and we have $\boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}, \boldsymbol{V} \in \mathbb{R}^{m\times r}$.

Proof To establish the result, it suffices to show that

$$\mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\boldsymbol{Y}\big) = \mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Y}\big).$$

Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be a singular value decomposition of $\boldsymbol{X}$ where $r = \mathrm{rank}(\boldsymbol{X})$ and $\boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}, \boldsymbol{V} \in \mathbb{R}^{m\times r}$. Observe that

$$\begin{aligned}
\mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T\boldsymbol{Y}\big) &= \mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^\dagger\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{Y}\big)\\
&= \mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T)^\dagger\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{Y}\big)\\
&= \mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{Y}\big)\\
&= \mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{Y}\big)\\
&= \mathrm{Tr}\big(\boldsymbol{Y}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Y}\big),
\end{aligned}$$

where we have repeatedly invoked the property that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_r$. ■

In light of Proposition 3, evaluating $f(\boldsymbol{X})$ for feasible solutions requires still requires $O(|\Omega|)$ operations for the first term, but the second term can be evaluated using $O(kn(m+d))$ operations (performing a truncated singular value decomposition is $O(knm)$ and computing the products involving $\boldsymbol{Y}$ is $O(knd)$) and the third term can be evaluated using $O(knm)$ operations (by performing a truncated singular value decomposition) for an overall complexity of $O(kn(m+d))$. This is significantly less expensive than the $O(m^2n + m^3 + n^2d)$ complexity of naive direct evaluation of $f(\boldsymbol{X})$ introduced previously.

4. An Exact Mixed-Projection Formulation

In this section, we reformulate (1) as a mixed-projection optimization problem and further reduce the dimension of the resulting problem in a commonly studied manner by parameterizing $\boldsymbol{X}$ as the matrix product of two low dimensional matrices. Thereafter, we illustrate how to employ the matrix generalization of the perspective relaxation (Günlük and Linderoth, 2012; Bertsimas et al., 2022, 2023c,a) to construct a convex relaxation of (1).

We first note that given the result of Section 3.2, we can rewrite (1) as an optimization problem only over $\boldsymbol{X}$ as follows:

$$\begin{aligned}
\min_{\boldsymbol{X}\in\mathbb{R}^{n\times m}} \quad & \sum_{(i,j)\in\Omega} (X_{ij} - A_{ij})^2 + \lambda\mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T)\boldsymbol{Y}\big) + \gamma\|\boldsymbol{X}\|_\star \\
\text{s.t.} \quad & \mathrm{rank}(\boldsymbol{X}) \leq k.
\end{aligned} \tag{5}$$

Observe that the matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^\dagger\boldsymbol{X}^T$ is the linear transformation that projects vectors onto the subspace spanned by the columns of the matrix $\boldsymbol{X}$. Drawing on ideas presented in Bertsimas et al. (2022, 2023c,a), we introduce an orthogonal projection matrix $\boldsymbol{P} \in \mathcal{P}_k$ to model the column space of $\boldsymbol{X}$ where $\mathcal{P}_\eta = \{\boldsymbol{P} \in \mathcal{S}^n : \boldsymbol{P}^2 = \boldsymbol{P}, \mathrm{tr}(\boldsymbol{P}) \leq \eta\}$ for $\eta \geq 0$. We can express the desired relationship between $\boldsymbol{P}$ and $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}$ since projecting a matrix onto its own column space leaves the matrix unchanged. This gives the following reformulation of (1):

$$\begin{aligned} \min_{\boldsymbol{X}\in\mathbb{R}^{n\times m}, \boldsymbol{P}\in\mathbb{R}^{n\times n}} \quad & \sum_{(i,j)\in\Omega} (X_{ij} - A_{ij})^2 + \lambda \mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}\big) + \gamma\|\boldsymbol{X}\|_\star \\ \text{s.t.} \quad & (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}_{n\times m},\ \boldsymbol{P} \in \mathcal{P}_{\min(k,\mathrm{rank}(\boldsymbol{X}))}. \end{aligned} \tag{6}$$

Observe that the matrix pseudo-inverse term has been eliminated from the objective function, however we have introduced the bilinear constraint $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}$ which is non convex in the optimization variables as well as the non convex constraint $\boldsymbol{P} \in \mathcal{P}_{\min(k,\mathrm{rank}(\boldsymbol{X}))}$. We now have the following result:

Proposition 4 *Problem* (6) *is a valid reformulation of* (5).

Proof We show that given a feasible solution to (6), we can construct a feasible solution to (5) that achieves the same objective value and vice versa.

Consider an arbitrary feasible solution $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ to (6). Since $\bar{\boldsymbol{P}}\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{P}} \in \mathcal{P}_{\min(k,\mathrm{rank}(\bar{\boldsymbol{X}}))}$, we have $\mathrm{rank}(\bar{\boldsymbol{X}}) \leq k$. We claim that $\bar{\boldsymbol{X}}$ achieves the same objective value in (5) as $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ achieves in (6). To show this, it suffices to illustrate that for all $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ feasible to (6) we have $H(\bar{\boldsymbol{X}}) \coloneqq \bar{\boldsymbol{X}}(\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}})^\dagger\bar{\boldsymbol{X}}^T = \bar{\boldsymbol{P}}$. The matrix $\bar{\boldsymbol{P}}$ is an orthogonal projection matrix since it is symmetric and satisfies $\bar{\boldsymbol{P}}^2 = \bar{\boldsymbol{P}}$. Moreover, since $\mathrm{rank}(\bar{\boldsymbol{P}}) = \mathrm{rank}(\bar{\boldsymbol{X}})$ and $\bar{\boldsymbol{P}}\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}$ we know that $\bar{\boldsymbol{P}}$ is an orthogonal projection onto the subspace spanned by the columns of $\bar{\boldsymbol{X}}$. Similarly, it can easily be verified that $H(\bar{\boldsymbol{X}})$ is symmetric and satisfies $H(\bar{\boldsymbol{X}})^2 = H(\bar{\boldsymbol{X}})$, $\mathrm{rank}(H(\bar{\boldsymbol{X}})) = \mathrm{rank}(\bar{\boldsymbol{X}})$ and $H(\bar{\boldsymbol{X}})\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}$. Thus, $H(\bar{\boldsymbol{X}})$ is also an orthogonal projection matrix onto the subspace spanned by the columns of $\bar{\boldsymbol{X}}$. To conclude, we invoke the property that given a subspace $\mathcal{V} \subset \mathbb{R}^n$ the orthogonal projection onto $\mathcal{V}$ is uniquely defined. To see this, suppose $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are two orthogonal projections onto $\mathcal{V}$. Let $l = \dim(\mathcal{V})$. Let $\{\boldsymbol{e}_i\}_{i=1}^l$ be an orthogonal basis for $\mathcal{V}$ and let $\{\boldsymbol{e}_i\}_{l+1}^n$ be an orthogonal basis for $\mathcal{V}^\perp$. Since $\boldsymbol{P}_1$ is an orthogonal projection onto $\mathcal{V}$, we have $\boldsymbol{P}_1\boldsymbol{e}_i = \boldsymbol{e}_i$ for all $1 \leq i \leq l$ and $\boldsymbol{P}_1\boldsymbol{e}_i = \boldsymbol{0}_n$ for all $l+1 \leq i \leq n$. However, the same must hold for $\boldsymbol{P}_2$ which implies that $\boldsymbol{P}_1 = \boldsymbol{P}_2$.

Consider an arbitrary feasible solution $\bar{\boldsymbol{X}}$ to (5). Let $r = \mathrm{rank}(\bar{\boldsymbol{X}})$ and $\bar{\boldsymbol{X}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^T$ be a singular value decomposition of $\bar{\boldsymbol{X}}$ where we have $\bar{\boldsymbol{U}} \in \mathbb{R}^{n\times r}, \bar{\boldsymbol{\Sigma}} \in \mathbb{R}^{r\times r}, \bar{\boldsymbol{V}} \in \mathbb{R}^{m\times r}$. Define $\bar{\boldsymbol{P}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T$. By construction, we have $\bar{\boldsymbol{P}} \in \mathcal{P}_{\min(k,\mathrm{rank}(\bar{\boldsymbol{X}}))}$ since $r \leq k$. Moreover, it is easy to verify that

$$\bar{\boldsymbol{P}}\bar{\boldsymbol{X}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^T = \bar{\boldsymbol{U}}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{V}}^T = \bar{\boldsymbol{X}},$$

where we have used the property $\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}} = \boldsymbol{I}_r$. Finally, Proposition 3 immediately implies that $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ achieves the same objective in (6) as $\bar{\boldsymbol{X}}$ achieves in (5). This completes the proof. ■

Optimizing explicitly over the space of $n \times m$ matrices can rapidly become prohibitively

costly in terms of runtime and memory requirements. Accordingly, we adopt the common approach of factorizing $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ as $\boldsymbol{U}\boldsymbol{V}^T$ for $\boldsymbol{U} \in \mathbb{R}^{n\times k}, \boldsymbol{V} \in \mathbb{R}^{m\times k}$. This leads to the following formulation:

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times k},\boldsymbol{V}\in\mathbb{R}^{m\times k},\\ \boldsymbol{P}\in\mathbb{R}^{n\times n}}} \quad & \sum_{(i,j)\in\Omega} ((\boldsymbol{U}\boldsymbol{V}^T)_{ij} - A_{ij})^2 + \lambda \mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}\big) + \frac{\gamma}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) \\ \text{s.t.} \quad & (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{U} = \boldsymbol{0}_{n\times k},\ \boldsymbol{P} \in \mathcal{P}_{\min(k,\mathrm{rank}(\boldsymbol{U}\boldsymbol{V}^T))}. \end{aligned} \tag{7}$$

Notice that we have replaced $n \times m$ optimization variables with $k \times (n + m)$ optimization variables, an often significant dimension reduction in practice. Attentive readers may object that though this is true, we have introduced n^2 decision variables through the introduction of the projection matrix variable $\boldsymbol{P}$ which nullifies any savings introduced through the factorization of $\boldsymbol{X}$. Note, however, that it is possible to factor any feasible projection matrix as $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{M}^T$ for some $\boldsymbol{M} \in \mathbb{R}^{n\times k}$. In Section 5, we leverage this fact so that the presence of the projection matrix incurs a cost of $n \times k$ additional variables rather than n^2 variables. We have the following result:

Proposition 5 *Problem* (7) *is a valid reformulation of* (6).

Proof We show that given a feasible solution to (7), we can construct a feasible solution to (6) that achieves the same or lesser objective value and vice versa.

Consider an arbitrary feasible solution $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ to (7). Let $\bar{\boldsymbol{X}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^T$. We will show that $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ is feasible to (6) and achieves the same or lesser objective as $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ does in (7). Feasibility of (7) implies that $\bar{\boldsymbol{P}} \in \mathcal{P}_{\min(k,\mathrm{rank}(\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^T))} = \mathcal{P}_{\min(k,\mathrm{rank}(\bar{\boldsymbol{X}}))}$ and also that

$$(\boldsymbol{I}_n - \bar{\boldsymbol{P}})\bar{\boldsymbol{X}} = (\boldsymbol{I}_n - \bar{\boldsymbol{P}})\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^T = \boldsymbol{0}_{n\times k}\bar{\boldsymbol{V}}^T = \boldsymbol{0}_{n\times m},$$

thus the solution $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ is certainly feasible for (6). To see that $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ achieves the same or lesser objective value, it suffices to argue that $\|\bar{\boldsymbol{X}}\|_\star \leq \frac{1}{2}(\|\bar{\boldsymbol{U}}\|_F^2 + \|\bar{\boldsymbol{V}}\|_F^2)$. This follows immediately from the following lemma established by Mazumder et al. (2010) (see Appendix A.5 in their paper for a proof):

Lemma 6 *For any matrix $\boldsymbol{Z}$, the following holds:*

$$\|\boldsymbol{Z}\|_\star = \min_{\boldsymbol{U},\boldsymbol{V}:\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{V}^T} \frac{1}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2).$$

If $\mathrm{rank}(\boldsymbol{Z}) = k \leq \min(m, n)$, then the minimum above is attained at a factor decomposition $\boldsymbol{U}_{n\times k}\boldsymbol{V}_{m\times k}^T$. Letting $\boldsymbol{Z}_{n\times m} = \boldsymbol{L}_{n\times k}\boldsymbol{\Sigma}_{k\times k}\boldsymbol{R}_{m\times k}^T$ denote a singular value decomposition of $\boldsymbol{Z}$, the minimum above is attained at $\boldsymbol{U}_{n\times k} = \boldsymbol{L}_{n\times k}\boldsymbol{\Sigma}_{k\times k}^{\frac{1}{2}}, \boldsymbol{V}_{m\times k} = \boldsymbol{R}_{m\times k}\boldsymbol{\Sigma}_{k\times k}^{\frac{1}{2}}$.

Consider now an arbitrary feasible solution $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ to (6). Let $\bar{\boldsymbol{X}} = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{R}^T$ be a singular value decomposition of $\bar{\boldsymbol{X}}$ where $\boldsymbol{L} \in \mathbb{R}^{n\times k}, \boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}, \boldsymbol{R} \in \mathbb{R}^{m\times k}$ and define $\bar{\boldsymbol{U}} = \boldsymbol{L}\boldsymbol{\Sigma}^{\frac{1}{2}}, \bar{\boldsymbol{V}} = \boldsymbol{R}\boldsymbol{\Sigma}^{\frac{1}{2}}$. Feasibility of $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ in (6) implies that $\bar{\boldsymbol{P}} \in \mathcal{P}_{\min(k,\mathrm{rank}(\bar{\boldsymbol{X}}))} =$

$\mathcal{P}_{\min(k,\text{rank}(\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^T))}$. Moreover, since the columns of $\boldsymbol{L}$ form an orthogonal basis for the columns space of $\bar{\boldsymbol{X}}$, the condition $(\boldsymbol{I}_n - \bar{\boldsymbol{P}})\bar{\boldsymbol{X}} = \boldsymbol{0}_{n\times m}$ implies that

$$(\boldsymbol{I}_n - \bar{\boldsymbol{P}})\bar{\boldsymbol{U}} = (\boldsymbol{I}_n - \bar{\boldsymbol{P}})\boldsymbol{L}\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{0}_{n\times k}\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{0}_{n\times k}.$$

Thus, the solution $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ is feasible to (7). Moreover, by Lemma 6 we have $\frac{1}{2}(\|\bar{\boldsymbol{U}}\|_F^2 + \|\bar{\boldsymbol{V}}\|_F^2) = \|\bar{\boldsymbol{X}}\|_\star$ so $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ achieves the same objective in (7) as $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ does in (6). This completes the proof.

■

In the remainder of the paper, we will relax the constraint $\boldsymbol{P} \in \mathcal{P}_{\min(k,\text{rank}(\boldsymbol{U}\boldsymbol{V}^T))}$ to $\boldsymbol{P} \in \mathcal{P}_k$ and develop a scalable algorithm to obtain high quality feasible solutions. Explicitly, we consider the problem given by:

$$\begin{aligned} \min_{\substack{\boldsymbol{U}\in\mathbb{R}^{n\times k}, \boldsymbol{V}\in\mathbb{R}^{m\times k},\\ \boldsymbol{P}\in\mathbb{R}^{n\times n}}} \quad & \sum_{(i,j)\in\Omega} ((\boldsymbol{U}\boldsymbol{V}^T)_{ij} - A_{ij})^2 + \lambda \text{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}\big) + \frac{\gamma}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) \\ \text{s.t.} \quad & (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{U} = \boldsymbol{0}_{n\times k},\ \boldsymbol{P} \in \mathcal{P}_k. \end{aligned} \tag{8}$$

It is straightforward to see that the optimal value of (8) is no greater than the optimal value of (7). Unfortunately, the converse does not necessarily hold. To see why the optimal value of (8) can be strictly less than that of (7) in certain pathological cases, suppose we had $k = n = m$, $\Omega = \emptyset$. In this setting, letting $\bar{\boldsymbol{P}} = \boldsymbol{I}_n$, $\boldsymbol{U} = \boldsymbol{0}_{n\times k}$ and $\bar{\boldsymbol{V}} = \boldsymbol{0}_{m\times k}$, the solution $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ would be feasible to (8) and achieve an objective value of 0. However the optimal value of (7) would be strictly greater than 0 in this setting as long as $\boldsymbol{Y} \neq 0$. Although (8) is a relaxation of (1), we will see in Section 6 that the solutions we will obtain to (8) will be high quality solutions for (1), the main problem of interest.

4.1 A Positive Semidefinite Cone Relaxation

Convex relaxations are useful in non convex optimization primarily for two reasons. Firstly, given the objective value achieved by an arbitrary feasible solution, strong convex relaxations can be used to upperbound the worst case suboptimality of said solution. Secondly, convex relaxations can often be used as building blocks for global optimization procedures. In this section, we present a natural convex relaxation of (8) that leverages the matrix generalization of the perspective relaxation (Günlük and Linderoth, 2012; Bertsimas et al., 2022, 2023c,a).

Rather than working directly with (8), consider the equivalent formulation (6) with $\mathcal{P}_{\min(k,\text{rank}(\boldsymbol{X}))}$ replaced by $\mathcal{P}_k$. Before proceeding, we will assume knowledge of an upper bound $M \in \mathbb{R}_+$ on the spectral norm of an optimal $\boldsymbol{X}$ to (6). Tighter bounds M are desirable as they will lead to stronger convex relaxations of (6). We note that it is always possible to specify such an upper bound M without prior knowledge of an optimal solution to (6). To see this, note that setting $\boldsymbol{X} = \boldsymbol{0}_{n\times m}$ in (6) produces an objective value of $\sum_{(i,j)\in\Omega} A_{ij}^2 + \lambda\|\boldsymbol{Y}\|_F^2$. Thus, any $\boldsymbol{X}$ such that $\gamma\|\boldsymbol{X}\|_\star > \sum_{(i,j)\in\Omega} A_{ij}^2 + \lambda\|\boldsymbol{Y}\|_F^2$ cannot possibly be optimal to (6). Finally, since the nuclear norm is an upper bound on the

spectral norm of a matrix, we must have

$$\|\boldsymbol{X}\|_\sigma \leq \frac{\sum_{(i,j)\in\Omega} A_{ij}^2 + \lambda\|\boldsymbol{Y}\|_F^2}{\gamma},$$

for any matrix $\boldsymbol{X}$ that is optimal to (6). We can therefore take $M = \frac{\sum_{(i,j)\in\Omega} A_{ij}^2 + \lambda\|\boldsymbol{Y}\|_F^2}{\gamma}$.

Notice that the non convexity in (6) is captured entirely by the bilinear constraint $(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}_{n\times m}$ and the quadratic constraint $\boldsymbol{P}^2 = \boldsymbol{P}$. In keeping with the approach presented in Bertsimas et al. (2023c,a), we leverage the matrix perspective to convexify the bilinear term and solve over the convex hull of the set $\mathcal{P}_k$. Recalling that the nuclear norm is semidefinite representable, we have the following formulation:

$$\begin{aligned}
\min_{\substack{\boldsymbol{P},\boldsymbol{W}_1\in\mathbb{R}^{n\times n},\\ \boldsymbol{X}\in\mathbb{R}^{n\times m},\boldsymbol{W}_2\in\mathbb{R}^{m\times m}}} \quad & \sum_{(i,j)\in\Omega} (X_{ij} - A_{ij})^2 + \lambda \mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}\big) + \frac{\gamma}{2}(\mathrm{Tr}(\boldsymbol{W}_1) + \mathrm{Tr}(\boldsymbol{W}_2)) \\
\text{s.t.} \quad & \boldsymbol{I}_n \succeq \boldsymbol{P} \succeq 0,\ \mathrm{Tr}(\boldsymbol{P}) \leq k, \\
& \begin{pmatrix} M\boldsymbol{P} & \boldsymbol{X} \\ \boldsymbol{X}^T & M\boldsymbol{I}_m \end{pmatrix} \succeq 0,\ \begin{pmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{W}_2 \end{pmatrix} \succeq 0.
\end{aligned} \tag{9}$$

We now have the following result:

Proposition 7 *Problem* (9) *is a valid convex relaxation of* (8).

Proof Problem (9) is clearly a convex optimization problem. We will show that the optimal value of (9) is a lower bound on the optimal value of (8) by showing that given any optimal solution to (8), we can construct a feasible solution to (9) that achieves the same objective value.

Consider any optimal solution $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ to (8). From the proof of Proposition 5, we know that the solution $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ where $\bar{\boldsymbol{X}} = \bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^T$ is feasible to (6) (where we replace the constraint $\boldsymbol{P} \in \mathcal{P}_{\min(k,\mathrm{rank}(\boldsymbol{U}\boldsymbol{V}^T))}$ with $\boldsymbol{P} \in \mathcal{P}_k$) and must also be optimal. Let $\bar{\boldsymbol{X}} = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{R}^T$ be a singular value decomposition of $\bar{\boldsymbol{X}}$ with $\boldsymbol{L} \in \mathbb{R}^{n\times k}, \boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{R} \in \mathbb{R}^{m\times k}$. Let $\bar{\boldsymbol{W}}_1 = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^T$ and $\bar{\boldsymbol{W}}_2 = \boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^T$. We claim that $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}}, \bar{\boldsymbol{W}}_1, \bar{\boldsymbol{W}}_2)$ is feasible to (9) and achieves the same objective value as $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ does in (6).

From the feasibility of $\bar{\boldsymbol{P}}$ in (8), we know that $\bar{\boldsymbol{P}} \in \mathcal{P}_k$ which implies $\boldsymbol{I}_n \succeq \bar{\boldsymbol{P}} \succeq 0$ and $\mathrm{Tr}(\bar{\boldsymbol{P}}) \leq k$. By the generalized Schur complement lemma (see Boyd et al. (1994), Equation 2.41), we know that

$$\begin{pmatrix} M\bar{\boldsymbol{P}} & \bar{\boldsymbol{X}} \\ \bar{\boldsymbol{X}}^T & M\boldsymbol{I}_m \end{pmatrix} \succeq 0 \iff M\boldsymbol{I}_m \succeq 0, \text{ and } M\boldsymbol{I}_m - \bar{\boldsymbol{X}}^T(M\bar{\boldsymbol{P}})^\dagger\bar{\boldsymbol{X}} \succeq 0.$$

We trivially have $M\boldsymbol{I}_m \succeq 0$. To see that the second condition holds, note that since $\bar{\boldsymbol{P}}$ is a projection matrix and $\bar{\boldsymbol{P}}\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}$, we have $\bar{\boldsymbol{X}}^T(M\bar{\boldsymbol{P}})^\dagger\bar{\boldsymbol{X}} = \frac{1}{M}\bar{\boldsymbol{X}}^T\bar{\boldsymbol{P}}\bar{\boldsymbol{X}} = \frac{1}{M}\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}$. Furthermore, since $\bar{\boldsymbol{X}}$ is optimal to (6), we have $\|\bar{\boldsymbol{X}}\|_\sigma \leq M$ by assumption. Thus, we have

$$\|\bar{\boldsymbol{X}}\|_\sigma \leq M \implies \|\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}\|_\sigma \leq M^2 \implies M^2\boldsymbol{I}_m \succeq \bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}} \implies M\boldsymbol{I}_m \succeq \frac{1}{M}\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}.$$

Finally, observe that

$$\begin{pmatrix} \bar{\boldsymbol{W}}_1 & \bar{\boldsymbol{X}} \\ \bar{\boldsymbol{X}}^T & \bar{\boldsymbol{W}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^T & \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{R}^T \\ \boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{L}^T & \boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{L} \\ \boldsymbol{R} \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} \boldsymbol{L} \\ \boldsymbol{R} \end{pmatrix}^T.$$

Since $\boldsymbol{\Sigma}$ is a diagonal matrix with non negative entries, the matrix $\begin{pmatrix} \bar{\boldsymbol{W}}_1 & \bar{\boldsymbol{X}} \\ \bar{\boldsymbol{X}}^T & \bar{\boldsymbol{W}}_2 \end{pmatrix}$ is certainly positive semidefinite. Thus we have shown that $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}}, \bar{\boldsymbol{W}}_1, \bar{\boldsymbol{W}}_2)$ is indeed feasible to (9). To conclude the proof, we note that

$$\begin{aligned} \frac{\gamma}{2}(\mathrm{Tr}(\bar{\boldsymbol{W}}_1) + \mathrm{Tr}(\bar{\boldsymbol{W}}_2)) &= \frac{\gamma}{2}(\mathrm{Tr}(\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^T) + \mathrm{Tr}(\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^T)) = \frac{\gamma}{2}(\mathrm{Tr}(\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{\Sigma}) + \mathrm{Tr}(\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{\Sigma})) \\ &= \frac{\gamma}{2}(\mathrm{Tr}(\boldsymbol{\Sigma}) + \mathrm{Tr}(\boldsymbol{\Sigma})) = \gamma\|\bar{\boldsymbol{X}}\|_\star, \end{aligned}$$

thus $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}}, \bar{\boldsymbol{W}}_1, \bar{\boldsymbol{W}}_2)$ achieves the same objective value in (9) as $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{P}})$ achieves in (6). ■

In general, an optimal solution to (9) will have $\boldsymbol{P} \notin \mathcal{P}_k$. We briefly note that to obtain a stronger convex relaxation, one could leverage eigenvector disjunctions (Bertsimas et al., 2023b; Saxena et al., 2010) to iteratively cut off solutions to (9) with $\boldsymbol{P} \notin \mathcal{P}_k$ and form increasingly tighter disjunctive approximations to the set $\mathcal{P}_k$.

5. Mixed-Projection ADMM

In this section, we present an alternating direction method of multipliers (ADMM) algorithm that is scalable and obtains high quality solutions for (8). Rather than forming the augmented lagrangian directly for (8), we first modify our problem formulation by introducing a dummy variable $\boldsymbol{Z} \in \mathbb{R}^{n\times k}$ that is an identical copy of $\boldsymbol{U}$. Additionally, rather than directly enforcing the constraint $\boldsymbol{P} \in \mathcal{P}_k$, we introduce an indicator function penalty $\mathbb{I}_{\mathcal{P}_k}(\boldsymbol{P})$ into the objective function where $\mathbb{I}_{\mathcal{X}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{X}$, otherwise $\mathbb{I}_{\mathcal{X}}(\boldsymbol{x}) = \infty$. Explicitly, we consider the following problem:

$$\begin{aligned} \min_{\substack{\boldsymbol{U},\boldsymbol{Z}\in\mathbb{R}^{n\times k},\\ \boldsymbol{V}\in\mathbb{R}^{m\times k},\\ \boldsymbol{P}\in\mathbb{R}^{n\times n}}} \quad & \sum_{(i,j)\in\Omega} ((\boldsymbol{U}\boldsymbol{V}^T)_{ij} - A_{ij})^2 + \lambda\mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}\big) + \frac{\gamma}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) + \mathbb{I}_{\mathcal{P}_k}(\boldsymbol{P}) \\ \text{s.t.} \quad & (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{U} = \boldsymbol{0}_{n\times k},\ \boldsymbol{U} - \boldsymbol{Z} = \boldsymbol{0}_{n\times k}. \end{aligned} \tag{10}$$

It is trivial to see that (10) is equivalent to (8). We will see in this section that working with formulation (10) leads to an ADMM algorithm with favorable decomposition properties. Introducing dual variables $\boldsymbol{\Phi}, \boldsymbol{\Psi} \in \mathbb{R}^{n\times k}$ for the constraints $(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{U} = \boldsymbol{0}_{n\times k}$ and $\boldsymbol{U} - \boldsymbol{Z} = \boldsymbol{0}_{n\times k}$ respectively, the augmented lagrangian $\mathcal{L}$ for (10) is given by:

$$\begin{aligned} \mathcal{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) = & \sum_{(i,j)\in\Omega} ((\boldsymbol{U}\boldsymbol{V}^T)_{ij} - A_{ij})^2 + \lambda\mathrm{Tr}\big(\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}\big) \\ & + \frac{\gamma}{2}(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) + \mathbb{I}_{\mathcal{P}_k}(\boldsymbol{P}) + \mathrm{Tr}(\boldsymbol{\Phi}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}) \\ & + \mathrm{Tr}(\boldsymbol{\Psi}^T(\boldsymbol{Z} - \boldsymbol{U})) + \frac{\rho_1}{2}\|(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}\|_F^2 + \frac{\rho_2}{2}\|\boldsymbol{Z} - \boldsymbol{U}\|_F^2, \end{aligned} \tag{11}$$

where $\rho_1, \rho_2 > 0$ are non-negative penalty parameters. In what follows, we show that performing a partial minimization of the augmented lagrangian (11) over each of the primal variables $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}$ yields a subproblem that can be solved efficiently. We present each subproblem and investigate the complexity of computing the corresponding subproblem solutions.

5.1 Subproblem in $\boldsymbol{U}$

First, suppose we fix variables $\boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ and seek to minimize $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ over $\boldsymbol{U}$. Eliminating terms that do not depend on $\boldsymbol{U}$, the resulting subproblem is given by

$$\min_{\boldsymbol{U} \in \mathbb{R}^{n \times k}} \sum_{(i,j) \in \Omega} ((\boldsymbol{U}\boldsymbol{V}^T)_{ij} - A_{ij})^2 + \frac{\gamma}{2}\|\boldsymbol{U}\|_F^2 - \mathrm{Tr}(\boldsymbol{\Psi}^T\boldsymbol{U}) + \frac{\rho_2}{2}\|\boldsymbol{Z} - \boldsymbol{U}\|_F^2. \tag{12}$$

We now have the following result:

Proposition 8 *The optimal solution $\bar{\boldsymbol{U}}$ for (12) is given by*

$$\bar{U}_{i,\star} = [2\boldsymbol{V}^T\boldsymbol{W}_i\boldsymbol{V} + (\gamma + \rho_2)\boldsymbol{I}_k]^{-1}[2\boldsymbol{V}^T\boldsymbol{W}_i A_{i,\star} + \Psi_{i,\star} + \rho_2 Z_{i,\star}], \tag{13}$$

for each $i \in \{1, \ldots, n\}$ where each $\boldsymbol{W}_i \in \mathbb{R}^{m \times m}$ is a diagonal matrix satisfying $(\boldsymbol{W}_i)_{jj} = 1$ if $(i,j) \in \Omega$, otherwise $(\boldsymbol{W}_i)_{jj} = 0$. Here, the column vectors $\bar{U}_{i,\star} \in \mathbb{R}^k, A_{i,\star} \in \mathbb{R}^m, \Psi_{i,\star} \in \mathbb{R}^k, Z_{i,\star} \in \mathbb{R}^k$ denote the i^{th} row of the matrices $\bar{\boldsymbol{U}}, \boldsymbol{A}, \boldsymbol{\Psi}, \boldsymbol{Z}$ respectively, where the unknown entries of $\boldsymbol{A}$ are taken to be 0.

Proof Let $f(\boldsymbol{U})$ denote the objective function of (12). With $\{\boldsymbol{W}_i\}_{i=1}^n$ defined as in Proposition 8, observe that we can write $f(\boldsymbol{U})$ as

$$\begin{aligned} f(\boldsymbol{U}) &= \sum_{i=1}^n \|\boldsymbol{W}_i(\boldsymbol{V}U_{i,\star} - A_{i,\star})\|_2^2 + \frac{\gamma}{2}\sum_{i=1}^n \|U_{i,\star}\|_2^2 - \sum_{i=1}^n \Psi_{i,\star}^T U_{i,\star} + \frac{\rho_2}{2}\sum_{i=1}^n \|Z_{i,\star} - U_{i,\star}\|_2^2 \\ &= \sum_{i=1}^n \left[\|\boldsymbol{W}_i(\boldsymbol{V}U_{i,\star} - A_{i,\star})\|_2^2 + \frac{\gamma}{2}\|U_{i,\star}\|_2^2 - \Psi_{i,\star}^T U_{i,\star} + \frac{\rho_2}{2}\|Z_{i,\star} - U_{i,\star}\|_2^2\right] = \sum_{i=1}^n g_i(\boldsymbol{U}), \end{aligned}$$

where we define $g_i(\boldsymbol{U}) = \|\boldsymbol{W}_i(\boldsymbol{V}U_{i,\star} - A_{i,\star})\|_2^2 + \frac{\gamma}{2}\|U_{i,\star}\|_2^2 - \Psi_{i,\star}^T U_{i,\star} + \frac{\rho_2}{2}\|Z_{i,\star} - U_{i,\star}\|_2^2$. Thus, we have shown that $f(\boldsymbol{U})$ is separable over the rows of the matrix $\boldsymbol{U}$. Each function $g_i(\boldsymbol{U})$ is a (strongly) convex quadratic. Thus, we can minimize $g_i(\boldsymbol{U})$ by setting its gradient to 0. For any fixed row $i \in \{1, \ldots, n\}$, we can differentiate and obtain

$$\nabla_{\boldsymbol{U}_{i,\star}} g_i(\boldsymbol{U}) = 2\boldsymbol{V}^T\boldsymbol{W}_i(\boldsymbol{V}U_{i,\star} - A_{i\star}) + \gamma U_{i,\star} - \Psi_{i,\star} - \rho_2(Z_{i,\star} - U_{i,\star}).$$

By equating the gradient $\nabla_{\boldsymbol{U}_{i,\star}} g_i(\boldsymbol{U})$ to 0 and rearranging, we obtain that the optimal vector $\bar{U}_{i,\star}$ is given by (13). This completes the proof. ■

Observe that since the matrix $\boldsymbol{V}^T\boldsymbol{W}_i\boldsymbol{V}$ is positive semidefinite and $\gamma + \rho_2 > 0$, the matrix inverse $[2\boldsymbol{V}^T\boldsymbol{W}_i\boldsymbol{V} + (\gamma + \rho_2)\boldsymbol{I}_k]^{-1}$ is well defined for all $i \in \{1, \ldots, n\}$. Computing the optimal solution to (12) requires computing n different $k \times k$ matrix inverses (where in general $k \ll \min\{m, n\}$). Computing a single $k \times k$ matrix inverse requires $O(k^3)$ time

and forming the matrix product $\boldsymbol{V}^T\boldsymbol{W}_i\boldsymbol{V}$ requires $O(k^2m)$ time for a given i. Thus, the complexity of computing the optimal solution for a single column is $O(k^3 + k^2m)$. Notice that each column of $\bar{\boldsymbol{U}}$ can be computed independently of the other columns. We leverage this observation by developing a multi-threaded implementation of the algorithm presented in this section. Letting w denote the number of compute threads available, computing the optimal solution $\bar{\boldsymbol{U}}$ of (12) requires $O\left(\frac{k^3n+k^2mn}{\min\{w,n\}}\right)$ time (the term $\min\{w, n\}$ in the denominator reflects that fact that increasing the number of available compute threads beyond the number of columns of $\bar{\boldsymbol{U}}$ does not yield additional reduction in compute complexity).

5.2 Subproblem in $\boldsymbol{V}$

Now, suppose we fix variables $\boldsymbol{U}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ and seek to minimize $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ over $\boldsymbol{V}$. Eliminating terms that do not depend on $\boldsymbol{V}$, the resulting subproblem is given by

$$\min_{\boldsymbol{V}\in\mathbb{R}^{m\times k}} \sum_{(i,j)\in\Omega} ((\boldsymbol{U}\boldsymbol{V}^T)_{ij} - A_{ij})^2 + \frac{\gamma}{2}\|\boldsymbol{V}\|_F^2. \tag{14}$$

We now have the following result:

Proposition 9 *The optimal solution $\bar{\boldsymbol{V}}$ for (14) is given by*

$$\bar{V}_{j,\star} = [2\boldsymbol{U}^T\boldsymbol{W}_j\boldsymbol{U} + \gamma\boldsymbol{I}_k]^{-1}[2\boldsymbol{U}^T\boldsymbol{W}_jA_{\star,j}], \tag{15}$$

for each $j \in \{1, \ldots, m\}$ where each $\boldsymbol{W}_j \in \mathbb{R}^{n\times n}$ is a diagonal matrix satisfying $(\boldsymbol{W}_j)_{ii} = 1$ if $(i, j) \in \Omega$, otherwise $(\boldsymbol{W}_j)_{ii} = 0$. Here, the column vector $\bar{V}_{j,\star} \in \mathbb{R}^k$ denotes the j^{th} row of $\bar{\boldsymbol{V}}$ while the column vector $A_{\star,j} \in \mathbb{R}^n$ denotes the j^{th} column of $\boldsymbol{A}$ where the unknown entries of $\boldsymbol{A}$ are taken to be 0.

Proof This proof follows the proof of Proposition 8. Let $f(\boldsymbol{V})$ denote the objective function of (14). With $\{\boldsymbol{W}_j\}_{j=1}^m$ defined as in Proposition 9, observe that we can write $f(\boldsymbol{V})$ as

$$\begin{aligned} f(\boldsymbol{V}) &= \sum_{j=1}^m \|\boldsymbol{W}_j(\boldsymbol{U}V_{j,\star} - A_{\star,j})\|_2^2 + \frac{\gamma}{2}\sum_{j=1}^m \|V_{j,\star}\|_2^2 \\ &= \sum_{j=1}^m \left[\|\boldsymbol{W}_j(\boldsymbol{U}V_{j,\star} - A_{\star,j})\|_2^2 + \frac{\gamma}{2}\|V_{j,\star}\|_2^2\right] = \sum_{j=1}^m g_j(\boldsymbol{V}), \end{aligned}$$

where we define $g_j(\boldsymbol{V}) = \|\boldsymbol{W}_j(\boldsymbol{U}V_{j,\star} - A_{\star,j})\|_2^2 + \frac{\gamma}{2}\|V_{j,\star}\|_2^2$. Thus, we have shown that $f(\boldsymbol{V})$ is separable over the rows of the matrix $\boldsymbol{V}$. Each function $g_j(\boldsymbol{V})$ is a (strongly) convex quadratic. Thus, we can minimize $g_j(\boldsymbol{V})$ by setting its gradient to 0. For any fixed row $j \in \{1, \ldots, m\}$, we can differentiate and obtain

$$\nabla_{\boldsymbol{V}_{j,\star}} g_j(\boldsymbol{V}) = 2\boldsymbol{U}^T\boldsymbol{W}_j(\boldsymbol{U}V_{j,\star} - A_{\star,j}) + \gamma V_{j,\star}.$$

By equating the gradient $\nabla_{\boldsymbol{V}_{j,\star}} g_j(\boldsymbol{V})$ to 0 and rearranging, we obtain that the optimal vector $\bar{U}_{i,\star}$ is given by (15). This completes the proof. ■

Observe that since the matrix $\boldsymbol{U}^T\boldsymbol{W}_j\boldsymbol{U}$ is positive semidefinite and $\gamma > 0$, the matrix inverse $[2\boldsymbol{U}^T\boldsymbol{W}_j\boldsymbol{U} + \gamma\boldsymbol{I}_k]^{-1}$ is well defined for all $j \in \{1, \ldots, m\}$. Computing the optimal solution to (14) requires computing m different $k \times k$ matrix inverses. Forming the matrix product $\boldsymbol{U}^T\boldsymbol{W}_j\boldsymbol{U}$ requires $O(k^2n)$ time for a given j. Thus, the complexity of computing the optimal solution for a single column is $O(k^3 + k^2n)$. Notice that, similarly to the solution of (12), each column of $\bar{\boldsymbol{V}}$ can be computed independently of the other columns. As before, we leverage this observation in our multi-threaded implementation of the algorithm presented in this section. Letting w denote the number of compute threads available, computing the optimal solution $\bar{\boldsymbol{V}}$ of (14) requires $O\Big(\frac{k^3m+k^2mn}{\min\{w,m\}}\Big)$ time.

The optimal solution $\bar{\boldsymbol{V}}$ to (14) reveals that the frobenius norm regularization term on $\boldsymbol{V}$ in (8) (which emerges from the nuclear norm regularization term on $\boldsymbol{X}$ in (1)) has computational benefits. Indeed, if we had $\gamma = 0$, it is possible that the matrix $\boldsymbol{U}^T\boldsymbol{W}_j\boldsymbol{U}$ be singular at certain iterates of our ADMM algorithm, in which case the corresponding matrix inverse would be undefined. This observation is in keeping with several recent works in the statistics, machine learning and operations research literatures where the presence of a regularization penalty in the objective function yields improved out of sample performance as well as benefits in computational tractability (see for example Bertsimas et al. (2020, 2021, 2023c,a); Bertsimas and Johnson (2024)).

5.3 Subproblem in $\boldsymbol{P}$

Now, suppose we fix variables $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ and seek to minimize $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ over $\boldsymbol{P}$. Eliminating terms that do not depend on $\boldsymbol{P}$, the resulting subproblem is given by

$$\begin{aligned} \min_{\boldsymbol{P}\in\mathbb{S}^n_+} \quad & -\lambda\mathrm{Tr}(\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{Y}) - \mathrm{Tr}(\boldsymbol{\Phi}^T\boldsymbol{P}\boldsymbol{Z}) + \frac{\rho_1}{2}\|(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{P}^2 = \boldsymbol{P},\ \mathrm{Tr}(\boldsymbol{P}) \leq k. \end{aligned} \tag{16}$$

We now have the following result:

Proposition 10 *Let $\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T$ be a rank k truncated singular value decomposition for the matrix given by:*

$$\Big(\lambda\boldsymbol{Y}\boldsymbol{Y}^T + \frac{\rho_1}{2}\boldsymbol{Z}\boldsymbol{Z}^T + \frac{1}{2}(\boldsymbol{\Phi}\boldsymbol{Z}^T + \boldsymbol{Z}\boldsymbol{\Phi}^T)\Big),$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}, \boldsymbol{M} \in \mathbb{R}^{n\times k}, \boldsymbol{M}^T\boldsymbol{M} = \boldsymbol{I}_k$. The optimal solution $\bar{\boldsymbol{P}}$ for (16) is given by $\bar{\boldsymbol{P}} = \boldsymbol{M}\boldsymbol{M}^T$.

Proof Let $f(\boldsymbol{P})$ denote the objective function of (16). Observe that for any $\boldsymbol{P}$ that is feasible to (16), we can write $f(\boldsymbol{P})$ as:

$$\begin{aligned} f(\boldsymbol{P}) &= -\lambda\mathrm{Tr}(\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{Y}) - \mathrm{Tr}(\boldsymbol{\Phi}^T\boldsymbol{P}\boldsymbol{Z}) + \frac{\rho_1}{2}\|(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}\|_F^2 \\ &= -\lambda\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}) - \mathrm{Tr}(\boldsymbol{Z}\boldsymbol{\Phi}^T\boldsymbol{P}) + \frac{\rho_1}{2}\mathrm{Tr}(\boldsymbol{Z}\boldsymbol{Z}^T(\boldsymbol{I}_n - \boldsymbol{P})) \\ &= \frac{\rho_1}{2}\mathrm{Tr}(\boldsymbol{Z}\boldsymbol{Z}^T) - \langle\lambda\boldsymbol{Y}\boldsymbol{Y}^T + \boldsymbol{Z}\boldsymbol{\Phi}^T + \frac{\rho_1}{2}\boldsymbol{Z}\boldsymbol{Z}^T, \boldsymbol{P}\rangle. \end{aligned}$$

Thus, it is immediately clear that a solution will be optimal to (16) if and only if it is optimal to the problem given by:

$$\begin{aligned} \max_{\boldsymbol{P} \in \mathbb{S}_+^n} \quad & \langle \boldsymbol{C}, \boldsymbol{P} \rangle \\ \text{s.t.} \quad & \boldsymbol{P}^2 = \boldsymbol{P},\ \mathrm{Tr}(\boldsymbol{P}) \leq k, \end{aligned} \tag{17}$$

where we define the matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ as $\boldsymbol{C} = \lambda \boldsymbol{Y}\boldsymbol{Y}^T + \boldsymbol{Z}\boldsymbol{\Phi}^T + \frac{\rho_1}{2}\boldsymbol{Z}\boldsymbol{Z}^T$. Let $\bar{\boldsymbol{C}} = \frac{1}{2}(\boldsymbol{C} + \boldsymbol{C}^T)$ denote the symmetric part of $\boldsymbol{C}$. Observe that for any symmetric matrix $\boldsymbol{P}$, we can consider $\langle \bar{\boldsymbol{C}}, \boldsymbol{P} \rangle$ in place of $\langle \boldsymbol{C}, \boldsymbol{P} \rangle$ since we have

$$\begin{aligned} \langle \boldsymbol{C}, \boldsymbol{P} \rangle &= \sum_{i=1}^n \sum_{j=1}^n P_{ij} C_{ij} = \sum_{i=1}^n P_{ii} C_{ii} + \sum_{i=1}^{n-1} \sum_{j=i+1}^n P_{ij}(C_{ij} + C_{ji}) \\ &= \sum_{i=1}^n P_{ii} \bar{C}_{ii} + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n P_{ij} \bar{C}_{ij} = \sum_{i=1}^n \sum_{j=1}^n P_{ij} \bar{C}_{ij} = \langle \bar{\boldsymbol{C}}, \boldsymbol{P} \rangle. \end{aligned}$$

Let $\bar{\boldsymbol{C}} = \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T$ be a full singular value decomposition of $\bar{\boldsymbol{C}}$ with $\boldsymbol{M}, \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, $\boldsymbol{M}^T\boldsymbol{M} = \boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{I}_n$. The matrix $\boldsymbol{\Sigma}$ is the diagonal matrix of (ordered) singular values of $\bar{\boldsymbol{C}}$ and we let σ_i denote the i^{th} singular value. Any feasible matrix $\boldsymbol{P}$ to (17) can be written as $\boldsymbol{P} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L} \in \mathbb{R}^{n \times k}, \boldsymbol{L}^T\boldsymbol{L} = \boldsymbol{I}_k$. Thus, for any $\boldsymbol{P}$ feasible to (17) we express the objective value as:

$$\langle \bar{\boldsymbol{C}}, \boldsymbol{P} \rangle = \mathrm{Tr}(\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{L}\boldsymbol{L}^T) = \mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{M}^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{M})) = \sum_{i=1}^n \sigma_i \|(\boldsymbol{M}^T\boldsymbol{L})_{i,\star}\|_2^2.$$

Let $\boldsymbol{N} = \boldsymbol{M}^T\boldsymbol{L} \in \mathbb{R}^{n \times k}$. Note that we have $\boldsymbol{N}^T\boldsymbol{N} = \boldsymbol{L}^T\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{L} = \boldsymbol{L}^T\boldsymbol{L} = \boldsymbol{I}_k$, which implies that the columns of $\boldsymbol{N}$ are orthonormal. This immediately implies that we have $N_{i,\star}^T N_{i,\star} = \|(\boldsymbol{M}^T\boldsymbol{L})_{i,\star}\|_2^2 \leq 1$. Moreover, we have

$$\sum_{i=1}^n \|(\boldsymbol{M}^T\boldsymbol{L})_{i,\star}\|_2^2 = \sum_{i=1}^n N_{i,\star}^T N_{i,\star} = \sum_{i=1}^n \sum_{j=1}^k N_{ij}^2 = \sum_{j=1}^k \sum_{i=1}^n N_{ij}^2 = \sum_{j=1}^k 1 = k.$$

We can therefore upper bound the optimal objective value of (17) as

$$\langle \bar{\boldsymbol{C}}, \boldsymbol{P} \rangle = \sum_{i=1}^n \sigma_i \|(\boldsymbol{M}^T\boldsymbol{L})_{i,\star}\|_2^2 \leq \sum_{i=1}^k \sigma_i.$$

To conclude the proof, notice that by taking $\bar{\boldsymbol{P}} = \bar{\boldsymbol{M}}\bar{\boldsymbol{M}}^T$ where $\bar{\boldsymbol{M}} \in \mathbb{R}^{n \times k}$ is the matrix that consists of the first k columns of $\boldsymbol{M}$ we can achieve the upper bound on (17):

$$\langle \bar{\boldsymbol{C}}, \bar{\boldsymbol{P}} \rangle = \mathrm{Tr}(\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T\bar{\boldsymbol{M}}\bar{\boldsymbol{M}}^T) = \mathrm{Tr}(\bar{\boldsymbol{M}}^T\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T\bar{\boldsymbol{M}}) = \sum_{i=1}^k \sigma_i.$$

■

To compute the optimal solution of (16), we need to compute a rank k singular value decomposition of the matrix $\bar{\boldsymbol{C}} = \left(\lambda \boldsymbol{Y}\boldsymbol{Y}^T + \frac{\rho_1}{2}\boldsymbol{Z}\boldsymbol{Z}^T + \frac{1}{2}(\boldsymbol{\Phi}\boldsymbol{Z}^T + \boldsymbol{Z}\boldsymbol{\Phi}^T)\right)$ which requires $O(kn^2)$ time since $\bar{\boldsymbol{C}} \in \mathbb{R}^{n\times n}$. Moreover, explicitly forming the matrix $\bar{\boldsymbol{C}}$ in memory from its constituent matrices $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Phi}$ requires $O(n^2(d+k))$ operations. Thus, naively computing the optimal solution to (16) has complexity $O(n^2(d+k))$ where the bottleneck operation from a complexity standpoint is explicitly forming the matrix $\bar{\boldsymbol{C}}$.

Fortunately, it is possible to compute the optimal solution to (16) more efficiently. Observe that we can equivalently express the matrix $\bar{\boldsymbol{C}}$ as $\bar{\boldsymbol{C}} = \boldsymbol{F}_1\boldsymbol{F}_2^T$ where $\boldsymbol{F}_1, \boldsymbol{F}_2 \in \mathbb{R}^{n\times(d+3k)}$ are defined as

$$\begin{aligned}\boldsymbol{F}_1 &= \begin{pmatrix}\sqrt{\lambda}\boldsymbol{Y} & \sqrt{\frac{\rho_1}{2}}\boldsymbol{Z} & \sqrt{\frac{1}{2}}\boldsymbol{\Phi} & \sqrt{\frac{1}{2}}\boldsymbol{Z}\end{pmatrix},\\ \boldsymbol{F}_2 &= \begin{pmatrix}\sqrt{\lambda}\boldsymbol{Y} & \sqrt{\frac{\rho_1}{2}}\boldsymbol{Z} & \sqrt{\frac{1}{2}}\boldsymbol{Z} & \sqrt{\frac{1}{2}}\boldsymbol{\Phi}\end{pmatrix}.\end{aligned}$$

Computing a truncated singular value decomposition requires only computing repeated matrix vector products. Therefore, rather than explicitly forming the matrix $\bar{\boldsymbol{C}}$ in memory at a cost of $O(n^2(d+k))$ operations, in our implementation we design a custom matrix class where matrix vector products between $\bar{\boldsymbol{C}}$ and arbitrary vectors $\boldsymbol{x} \in \mathbb{R}^n$ are computed by first evaluating the matrix vector product $\boldsymbol{\nu} = \boldsymbol{F}_2^T\boldsymbol{x}$ and subsequently evaluating the matrix vector product $\bar{\boldsymbol{C}}\boldsymbol{x} = \boldsymbol{F}_1\boldsymbol{\nu}$. In so doing, we can evaluate matrix vector products $\bar{\boldsymbol{C}}\boldsymbol{x}$ in $O(n(d+k))$ time rather than $O(n^2)$ time (in general, we will have $d+k \ll n$). Computing a truncated singular value decomposition of $\bar{\boldsymbol{C}}$ with this methodology of evaluating matrix vector products requires only $O(k^2n + knd)$ operations. Thus, our custom matrix class implementation avoids needing to explicitly for $\bar{\boldsymbol{C}}$ in memory and allows the optimal solution to (16) to be computed in $O(k^2n + knd)$ time.

5.4 Subproblem in $\boldsymbol{Z}$

Now, suppose we fix variables $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{\Phi}, \boldsymbol{\Psi}$ and seek to minimize $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{P}, \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ over $\boldsymbol{Z}$. Eliminating terms that do not depend on $\boldsymbol{Z}$, the resulting subproblem is given by

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times k}} \operatorname{Tr}(\boldsymbol{\Phi}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}) + \operatorname{Tr}(\boldsymbol{\Psi}^T\boldsymbol{Z}) + \frac{\rho_1}{2}\|(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z}\|_F^2 + \frac{\rho_2}{2}\|\boldsymbol{Z} - \boldsymbol{U}\|_F^2. \tag{18}$$

We now have the following result:

Proposition 11 *The optimal solution $\bar{\boldsymbol{Z}}$ for (18) is given by*

$$\begin{aligned}\bar{\boldsymbol{Z}} &= \frac{1}{\rho_1+\rho_2}\left(\boldsymbol{I}_n + \frac{\rho_1}{\rho_2}\boldsymbol{P}\right)\left(\rho_2\boldsymbol{U} - (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\Phi} - \boldsymbol{\Psi}\right)\\ &= \frac{1}{\rho_1+\rho_2}\left(\rho_2\boldsymbol{U} - \boldsymbol{\Phi} + \boldsymbol{P}\boldsymbol{\Phi} - \boldsymbol{\Psi} + \rho_1\boldsymbol{P}\boldsymbol{U} - \frac{\rho_1}{\rho_2}\boldsymbol{P}\boldsymbol{\Psi}\right).\end{aligned} \tag{19}$$

Proof Let $f(\boldsymbol{Z})$ denote the objective function of (18). The function $f(\boldsymbol{Z})$ is a convex quadratic, thus it can be minimized by setting its gradient to 0. Differentiating $f(\boldsymbol{Z})$, we obtain:

$$\nabla_{\boldsymbol{Z}} f(\boldsymbol{Z}) = (\boldsymbol{I}_n - \boldsymbol{P})^T\boldsymbol{\Phi} + \boldsymbol{\Psi} + \rho_1(\boldsymbol{I}_n - \boldsymbol{P})^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z} + \rho_2(\boldsymbol{Z} - \boldsymbol{U}).$$

Moreover, for any matrix $\boldsymbol{P}$ for which the augmented lagrangian (11) takes finite value, we will have $\boldsymbol{P} \in \mathcal{P}_k$ which implies that $\boldsymbol{P}^T = \boldsymbol{P}$ and $\boldsymbol{P}^2 = \boldsymbol{P}$. We can therefore simplify $\nabla_{\boldsymbol{Z}} f(\boldsymbol{Z})$ as:

$$\nabla_{\boldsymbol{Z}} f(\boldsymbol{Z}) = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\Phi} + \boldsymbol{\Psi} + \rho_1(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Z} + \rho_2(\boldsymbol{Z} - \boldsymbol{U}).$$

By equating the gradient $\nabla_{\boldsymbol{Z}} f(\boldsymbol{Z})$ to 0 and rearranging, we obtain that the optimal matrix $\bar{\boldsymbol{Z}}$ is given by:

$$\bar{\boldsymbol{Z}} = \Big(\rho_1(\boldsymbol{I}_n - \boldsymbol{P}) + \rho_2 \boldsymbol{I}_n\Big)^{-1}\Big(\rho_2 \boldsymbol{U} - (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\Phi} - \boldsymbol{\Psi}\Big)$$

To conclude the proof, it remains to show that $\big(\rho_1(\boldsymbol{I}_n - \boldsymbol{P}) + \rho_2 \boldsymbol{I}_n\big)^{-1} = \frac{1}{\rho_1+\rho_2}\big(\boldsymbol{I}_n + \frac{\rho_1}{\rho_2}\boldsymbol{P}\big)$. Let $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{M}^T$ where $\boldsymbol{M} \in \mathbb{R}^{n\times k}, \boldsymbol{M}^T\boldsymbol{M} = \boldsymbol{I}_k$. Such a matrix $\boldsymbol{M}$ is guaranteed to exist for any $\boldsymbol{P} \in \mathcal{P}_k$. We have

$$\begin{aligned}
\rho_1(\boldsymbol{I}_n - \boldsymbol{P}) + \rho_2 \boldsymbol{I}_n &= \rho_1(\boldsymbol{I}_n - \boldsymbol{M}\boldsymbol{M}^T) + \rho_2 \boldsymbol{I}_n \\
&= (\rho_1 + \rho_2)\boldsymbol{I}_n + \boldsymbol{M}(-\rho_1 \boldsymbol{I}_n)\boldsymbol{M}^T \\
&= \frac{1}{\rho_1+\rho_2}\boldsymbol{I}_n - \frac{1}{(\rho_1+\rho_2)^2}\boldsymbol{M}\left(\frac{1}{\rho_1+\rho_2}\boldsymbol{M}^T\boldsymbol{M} - \frac{1}{\rho_1}\boldsymbol{I}_k\right)^{-1}\boldsymbol{M}^T \\
&= \frac{1}{\rho_1+\rho_2}\boldsymbol{I}_n - \frac{1}{(\rho_1+\rho_2)^2}\boldsymbol{M}\left(\frac{-\rho_1(\rho_1+\rho_2)}{\rho_2}\boldsymbol{I}_k\right)\boldsymbol{M}^T \\
&= \frac{1}{\rho_1+\rho_2}\Big(\boldsymbol{I}_n + \frac{\rho_1}{\rho_2}\boldsymbol{P}\Big),
\end{aligned}$$

where the third equality follows from the Woodbury matrix inversion lemma (see Petersen et al. (2008), Section 3.2.2). As a sanity check, one can verify that the product of $\big(\rho_1(\boldsymbol{I}_n - \boldsymbol{P}) + \rho_2 \boldsymbol{I}_n\big)$ and $\frac{1}{\rho_1+\rho_2}\big(\boldsymbol{I}_n + \frac{\rho_1}{\rho_2}\boldsymbol{P}\big)$ is indeed the n dimensional identity matrix. ■

Evaluating the optimal solution to (18) requires only matrix-matrix multiplications. Computing the products of $\boldsymbol{P}\boldsymbol{\Phi}, \boldsymbol{P}\boldsymbol{U}, \boldsymbol{P}\boldsymbol{\Psi}$ in the definition of $\bar{\boldsymbol{Z}}$ from (19) requires $O(kn^2)$ operations. Thus, the naive cost of forming $\bar{\boldsymbol{Z}}$ is $O(kn^2)$. However, notice that if we had a factored representation of the matrix $\boldsymbol{P}$ as $\boldsymbol{P} = \boldsymbol{M}\boldsymbol{M}^T$ with $\boldsymbol{M} \in \mathbb{R}^{n\times k}$, for any matrix $\boldsymbol{R} \in \mathbb{R}^{n\times k}$ we could compute matrix-matrix products $\boldsymbol{P}\boldsymbol{R}$ by first computing $\boldsymbol{S} = \boldsymbol{M}^T\boldsymbol{R}$ and thereafter computing $\boldsymbol{P}\boldsymbol{R} = \boldsymbol{M}\boldsymbol{S}$ for a total complexity of $O(k^2 n)$. One might object that this ignores the time required to compute such a matrix $\boldsymbol{M}$. However, observe that in computing a matrix $\boldsymbol{P}$ that is optimal to (16), we in fact must already generate such a matrix $\boldsymbol{M}$ (see proposition 10). In fact, in our implementation we never explicitly form a $n \times n$ matrix $\boldsymbol{P}$ as it suffices to only store a copy of its low rank factorization matrix $\boldsymbol{M}$. Thus, the optimal solution to (18) can be evaluated in $O(k^2 n)$ time.

5.5 An ADMM Algorithm

Having illustrated that the partial minimization of the lagrangian (11) across each of the primal variables (Problems (12), (14), (16), (18)) can be solved efficiently, we can now present the overall approach Algorithm 1.

We initialize primal iterates $\boldsymbol{U}_0 = \boldsymbol{Z}_0 = \boldsymbol{L}\boldsymbol{\Sigma}^{\frac{1}{2}}, \boldsymbol{P}_0 = \boldsymbol{L}\boldsymbol{L}^T, \boldsymbol{V}_0 = \boldsymbol{R}\boldsymbol{\Sigma}^{\frac{1}{2}}$ where $\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{R}$ denotes a rank k truncated singular value decomposition of $\boldsymbol{A}$ (the missing entries of $\boldsymbol{A}$

Algorithm 1: Mixed-Projection ADMM

Data: $n, m, k \in \mathbb{Z}^+, \Omega \subset [n] \times [m], \{A_{ij}\}_{(i,j)\in\Omega}, \lambda, \gamma \in \mathbb{R}^+$. Tolerance parameter $\epsilon > 0$. Maximum iteration parameter $T \in \mathbb{Z}^+$

Result: $(\bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}, \bar{\boldsymbol{P}})$ that is feasible to (8).

$(\boldsymbol{U}_0, \boldsymbol{P}_0, \boldsymbol{V}_0, \boldsymbol{Z}_0) \leftarrow (\boldsymbol{L}\boldsymbol{\Sigma}^{\frac{1}{2}}, \boldsymbol{L}\boldsymbol{L}^T, \boldsymbol{R}\boldsymbol{\Sigma}^{\frac{1}{2}}, \boldsymbol{L}\boldsymbol{\Sigma}^{\frac{1}{2}})$ where $\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{R}$ is a rank k truncated SVD of $\boldsymbol{A}$ and missing entries are filled in with 0;

$(\boldsymbol{\Phi}_0, \boldsymbol{\Psi}_0) \leftarrow (\boldsymbol{1}_{n\times k}, \boldsymbol{1}_{n\times k})$;

$t \leftarrow 0$;

while $t < T$ *and* $\max\{\|(\boldsymbol{I}_n - \boldsymbol{P}_t)\boldsymbol{Z}_t\|_F^2, \|\boldsymbol{Z}_t - \boldsymbol{U}_t\|_F^2\} > \epsilon$ **do**

- $(\boldsymbol{U}_{t+1}, \boldsymbol{P}_{t+1}) \leftarrow \operatorname{argmin}_{\boldsymbol{U},\boldsymbol{P}} \mathcal{L}(\boldsymbol{U}, \boldsymbol{V}_t, \boldsymbol{P}, \boldsymbol{Z}_t, \boldsymbol{\Phi}_t, \boldsymbol{\Psi}_t)$;
- $(\boldsymbol{V}_{t+1}, \boldsymbol{Z}_{t+1}) \leftarrow \operatorname{argmin}_{\boldsymbol{V},\boldsymbol{Z}} \mathcal{L}(\boldsymbol{U}_{t+1}, \boldsymbol{V}, \boldsymbol{P}_{t+1}, \boldsymbol{Z}, \boldsymbol{\Phi}_t, \boldsymbol{\Psi}_t)$;
- $\boldsymbol{\Phi}_{t+1} \leftarrow \boldsymbol{\Phi}_t + \rho_1(\boldsymbol{I} - \boldsymbol{P}_{t+1})\boldsymbol{Z}_{t+1}$;
- $\boldsymbol{\Psi}_{t+1} \leftarrow \boldsymbol{\Psi}_t + \rho_2(\boldsymbol{Z}_{t+1} - \boldsymbol{U}_{t+1})$;
- $t \leftarrow t + 1$;

end

return $(\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{P}_t)$

are filled in with 0s) and we initialize dual iterates $\boldsymbol{\Phi}_0 = \boldsymbol{\Psi}_0 = \boldsymbol{1}_{n\times k}$. Observe that the subproblems (12) and (16) can be solved simultaneously. Similarly, the subprobems (14) and (18) can be solved simultaneously. At each iteration of Algorithm 1, we first update the iterates $\boldsymbol{U}_{t+1}, \boldsymbol{P}_{t+1}$ by solving problems (12) and (16) with $(\boldsymbol{V}_t, \boldsymbol{Z}_t, \boldsymbol{\Phi}_t, \boldsymbol{\Psi}_t)$ fixed. Next, we update the iterates $\boldsymbol{V}_{t+1}, \boldsymbol{Z}_{t+1}$ by solving problems (14) and (18) with $(\boldsymbol{U}_{t+1}, \boldsymbol{P}_{t+1}, \boldsymbol{\Phi}_t, \boldsymbol{\Psi}_t)$ fixed. Finally, we update the dual iterates $\boldsymbol{\Phi}, \boldsymbol{\Psi}$ by taking a gradient ascent step. The gradients of the augmented lagrangian (11) with respect to $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are given by the primal residuals $(\boldsymbol{I}_n - \boldsymbol{P}_{t+1})\boldsymbol{Z}_{t+1}$ and $\boldsymbol{Z}_{t+1} - \boldsymbol{U}_{t+1}$ respectively. We use ρ_1 and ρ_2 respectively as the step size. We proceed until the squared norm of each primal residual is below a numerical tolerance parameter ϵ or until we reach an input maximum number of iterations T. We know have the following result:

Proposition 12 *Assume that the number of compute threads w is less than* $\min\{n, m\}$. *The per iteration complexity of Algorithm 1 is* $O\big(k^2 n + knd + \frac{k^3(n+m)+k^2nm}{w}\big)$.

Proof The result follows from the complexity analysis of problems (12), (14), (16) and (18). ■

6. Computational Results

We evaluate the performance of Algorithm 1 implemented in Julia 1.7.3. Throughout, we fix $\rho_1 = \rho_2 = 10$, set the maximum number of iterations $T = 20$ and set the number of compute threads $w = 24$. Note that given the novelty of Problem (1), there are no pre-existing specialized methods to benchmark against. Accordingly, we compare the performance of Algorithm 1 against well studied methods for the very closely related MC problem as well as a highly performant generic method for low rank matrix optimization

problems. The MC methods we consider are Fast-Impute (Bertsimas and Li, 2020), Soft-Impute (Mazumder et al., 2010) and Iterative-SVD (Troyanskaya et al., 2001) which we introduced formally in Section 2.1. We utilize the implementation of Fast-Impute made publicly available by (Bertsimas and Li, 2020) while we use the implementation of Soft-Impute and Iterative-SVD from the python package fancyimpute 0.7.0 (Rubinsteyn and Feldman, 2016). The matrix optimization method we consider is ScaledGD (scaled gradient descent) (Tong et al., 2021) which we introduced formally in Section 2.2.1 and implement ourselves. All experiments were performed using synthetic data on MIT's Supercloud Cluster (Reuther et al., 2018), which hosts Intel Xeon Platinum 8260 processors. To bridge the gap between theory and practice, we have made our code freely available on `GitHub` at `github.com/NicholasJohnson2020/LearningLowRankMatrices`.

To evaluate the performance of Algorithm 1, Fast-Impute, Soft-Impute, Iterative-SVD and ScaledGD on synthetic data, we consider the objective value achieved by a returned solution in (1), the ℓ_2 reconstruction error between a returned solution and the ground truth, the numerical rank of a returned solution and the execution time of each algorithm. Explicitly, let $\hat{\boldsymbol{X}} \in \mathbb{R}^{n \times m}$ denote the solution returned by a given method (where we define $\hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{V}}^T$ if the method outputs low rank factors $\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}$) and let $\boldsymbol{A}^{true} \in \mathbb{R}^{n \times m}$ denote the ground truth matrix. We define the the ℓ_2 reconstruction error of $\hat{\boldsymbol{X}}$ as

$$ERR_{\ell_2}(\hat{\boldsymbol{X}}) = \frac{\|\hat{\boldsymbol{X}} - \boldsymbol{A}^{true}\|_F^2}{\|\boldsymbol{A}^{true}\|_F^2}.$$

We compute the numerical rank of $\hat{\boldsymbol{X}}$ by calling the default rank function from the Julia LinearAlgebra package. We aim to answer the following questions:

1. How does the performance of Algorithm 1 compare to existing methods such as Fast-Impute, Soft-Impute, Iterative-SVD and ScaledGD on synthetic data?

2. How is the performance of Algorithm 1 affected by the number of rows n, the number of columns m, the dimension of the side information d and the underlying rank k of the ground truth?

3. Empircially, which subproblem solution update is the computational bottleneck of Algorithm 1?

6.1 Synthetic Data Generation

To generate synthetic data, we specify a number of rows $n \in \mathbb{Z}_+$, a number of columns $m \in \mathbb{Z}_+$, a desired rank $k \in \mathbb{Z}_+$ with $k < \min\{n, m\}$, the dimension of the side information $d \in \mathbb{Z}_+$, a fraction of missing values $\alpha \in (0, 1)$ and a noise parameter $\sigma \in \mathbb{R}_+$ that controls the signal to noise ratio. We sample matrices $\boldsymbol{U} \in \mathbb{R}^{n \times k}, \boldsymbol{V} \in \mathbb{R}^{m \times k}, \boldsymbol{\beta} \in \mathbb{R}^{m \times d}$ by drawing each entry $U_{ij}, V_{ij}, \beta_{ij}$ independently from the uniform distribution on the interval $[0, 1]$. Furthermore, we sample a noise matrix $\boldsymbol{N} \in \mathbb{R}^{n \times d}$ by drawing each entry N_{ij} independently from the univariate normal distribution with mean 0 and variance σ^2. We let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^T$ and we let $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{N}$. Lastly, we sample $\lfloor \alpha \cdot n \cdot m \rfloor$ indices uniformly at random from the collection $\mathcal{I} = \{(i, j) : 1 \leq i \leq n, 1 \leq j \leq m\}$ to be the set of missing indices, which we denote by Γ. The set of revealed entries can then be defined as $\Omega = \mathcal{I} \setminus \Gamma$. We

fix $\alpha = 0.9, \sigma = 2$ throughout our experiments and report numerical results for various different combinations of (n, m, d, k).

6.2 Sensitivity to Row Dimension

We present a comparison of Algorithm 1 with ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD as we vary the number of rows n. In these experiments, we fixed $m = 100, k = 5$, and $d = 150$ across all trials. We varied $n \in \{100, 200, 400, 800, 1000, 2000, 5000, 10000\}$ and we performed 20 trials for each value of n. For ScaledGD, we set the step size to be $\eta = \frac{1}{10\sigma_1(\boldsymbol{A})}$ where $\sigma_1(\boldsymbol{A})$ denotes the largest singular value of the input matrix $\boldsymbol{A}$ where we fill the unobserved entries with the value 0. Letting $f(\boldsymbol{U}_t, \boldsymbol{V}_t)$ denote the objective value achieved after iteration t of ScaledGD, we terminate ScaledGD when either $t > 1000$ or $\frac{f(\boldsymbol{U}_{t-1}, \boldsymbol{V}_{t-1}) - f(\boldsymbol{U}_t, \boldsymbol{V}_t)}{f(\boldsymbol{U}_{t-1}, \boldsymbol{V}_{t-1})} < 10^{-3}$. In words, we terminate ScaledGD after 1000 iterations or after the relative objective value improvement between two iterations is less than 0.1%.

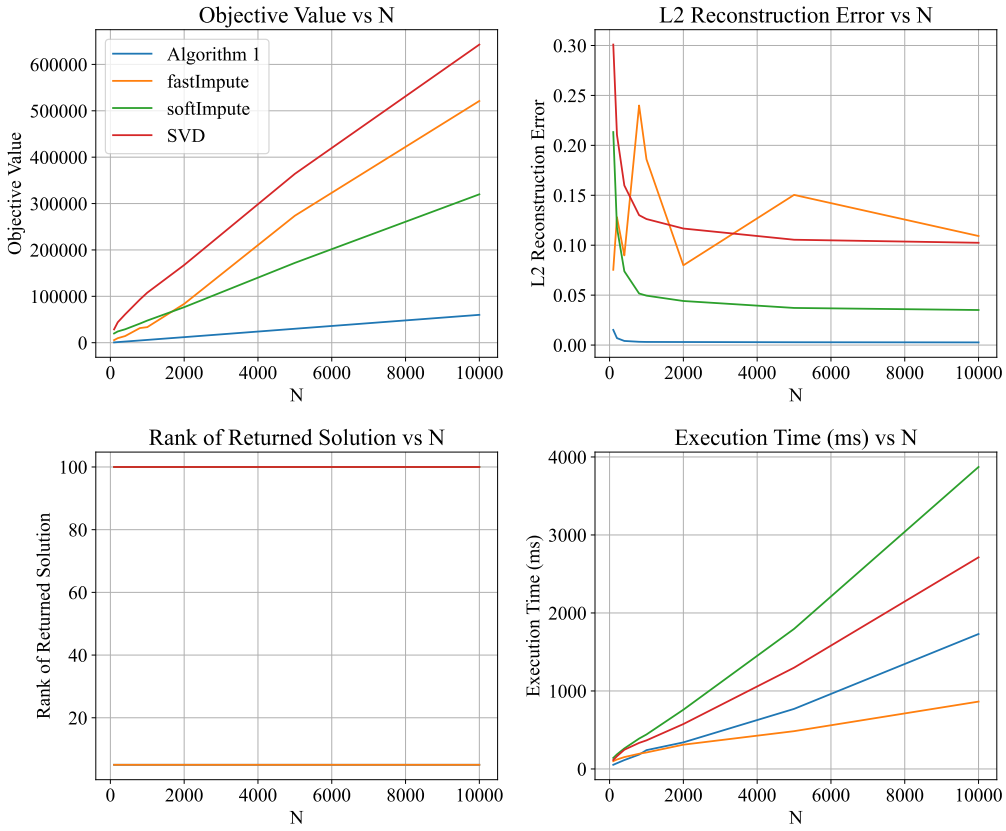

Figure 1: Objective value (top left), ℓ_2 reconstruction error (top right), fitted rank (bottom left) and execution time (bottom right) versus n with $m = 100, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

We report the objective value, ℓ_2 reconstruction error, fitted rank and execution time for Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Figure 1. We additionally report the objective value, reconstruction error and execution time for ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Tables 1, 2 and 3 of Appendix A. In Figure 2, we plot the average cumulative time spent solving subproblems (12), (14), (16), (18) during the execution of Algorithm 1 versus n. Our main findings from this set of experiments are:

1. Algorithm 1 systematically produces higher quality solutions than ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD (see Table 1), sometimes achieving an objective value that is an order of magnitude superior than the next best method. On average, Algorithm 1 outputs a solution whose objective value is 86% lesser than the objective value achieved by the best performing alternative method (Fast-Impute). We remind the reader that Fast-Impute, Soft-Impute and Iterative-SVD are methods designed for the generic MC problem and are not custom built to solve (1) so it should not come as a surprise that Algorithm 1 significantly outperforms these 3 methods in terms of objective value. ScaledGD however has explicit knowledge of the objective function of (1) along with its gradient, yet surprisingly produces the weakest average objective value across these experiments. We note that we use the default hyperparameters for ScaledGD recommended by the authors of this method (Tong et al., 2021). We observe that the objective value achieved by all methods increases linearly as the number of rows n increases.

2. In terms of ℓ_2 reconstruction error, Algorithm 1 again systematically produces solutions that are of higher quality than ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD (see Table 2), often achieving an error that is an order of magnitude superior than the next best method. On average, Algorithm 1 outputs a solution whose ℓ_2 reconstruction error is 92% lesser than the reconstruction error achieved by the best performing alternative method (Soft-Impute in all but one parameter configuration). This is especially noteworthy since Algorithm 1 is not designed explicitly with reconstruction error minimization as the objective, unlike Fast-Impute and Soft-Impute, and suggests that the side information $\boldsymbol{Y}$ is instrumental in recovering high quality low rank estimates of the partially observed data matrix.

3. We observe that the fitted rank of the solutions returned by Algorithm 1, ScaledGD and Fast-Impute always matched the specified target rank as would be expected, but surprisingly the solutions returned by Soft-Impute and Iterative-SVD were always of full rank despite the fact that these methods were provided with the target rank explicitly. This is potentially due to a numerical issues in the computation of the rank due to presence of extremely small singular values.

4. The runtime of Algorithm 1 is competitive with that of the other methods. The runtime of Algorithm 1 is less than of Soft-Impute and Iterative-SVD but greater than that of Fast-Impute. For experiments with $n \leq 2000$, Table 3 illustrates that ScaledGD was the method with the fastest execution time (however as previously mentioned the returned solutions were of low quality). The runtime of Algorithm 1, Fast-Impute, Soft-Impute and iterate SVD appear to grow linearly with n.

5. Figure 2 illustrates that the computation of the solution for (12) is the computational bottleneck in the execution of Algorithm 1 in this set of experiments, followed next by the computation of the solution for (16). Empirically, we observe that the solution time of (12), (14), (16) and (18) appear to scale linearly with the number of rows n. This observation is consistent with the computational complexities derived for each subproblem of Algorithm 1 in Section 5.

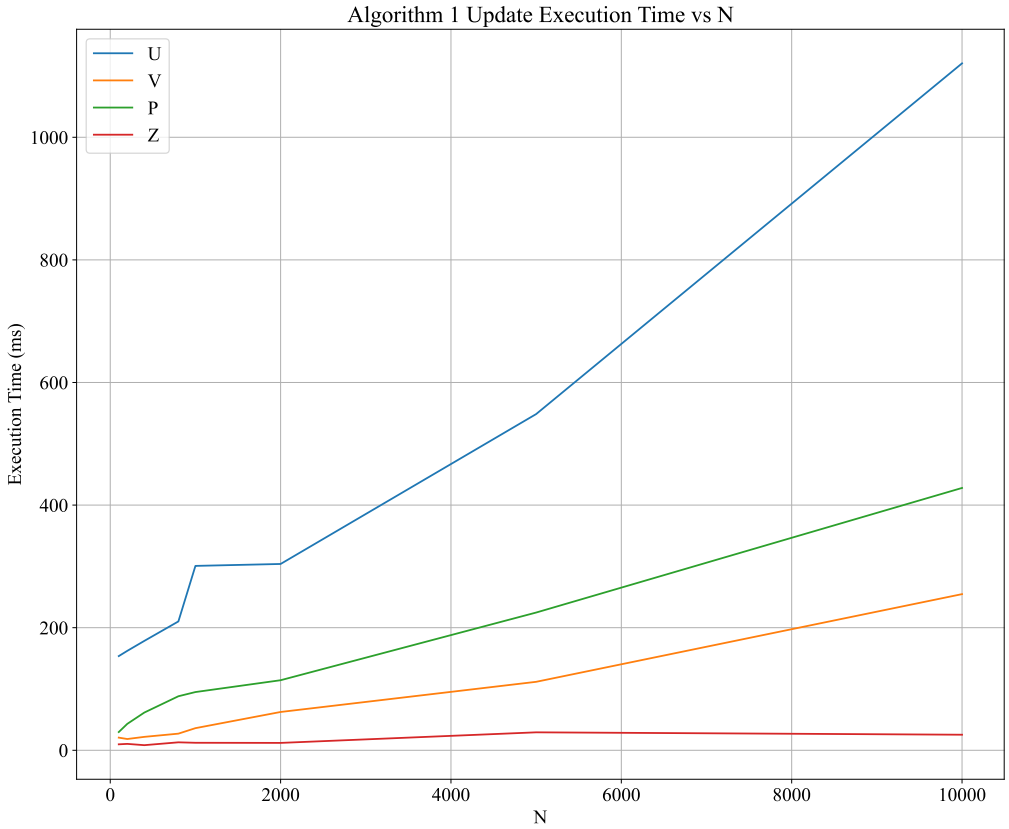

Figure 2: Cumulative time spent solving each subproblem of Algorithm 1 versus n with $m = 100, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

6.3 Sensitivity to Column Dimension

Here, we present a comparison of Algorithm 1 with ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD as we vary the number of columns m. We fixed $n = 1000, k = 5$, and $d = 150$ across all trials. We varied $m \in \{100, 200, 400, 800, 1000, 2000, 5000, 10000\}$ and we performed 20 trials for each value of m.

We report the objective value, ℓ_2 reconstruction error, fitted rank and execution time for Algorithm 1, Fast-Impute and Soft-Impute in Figure 3. We additionally report the

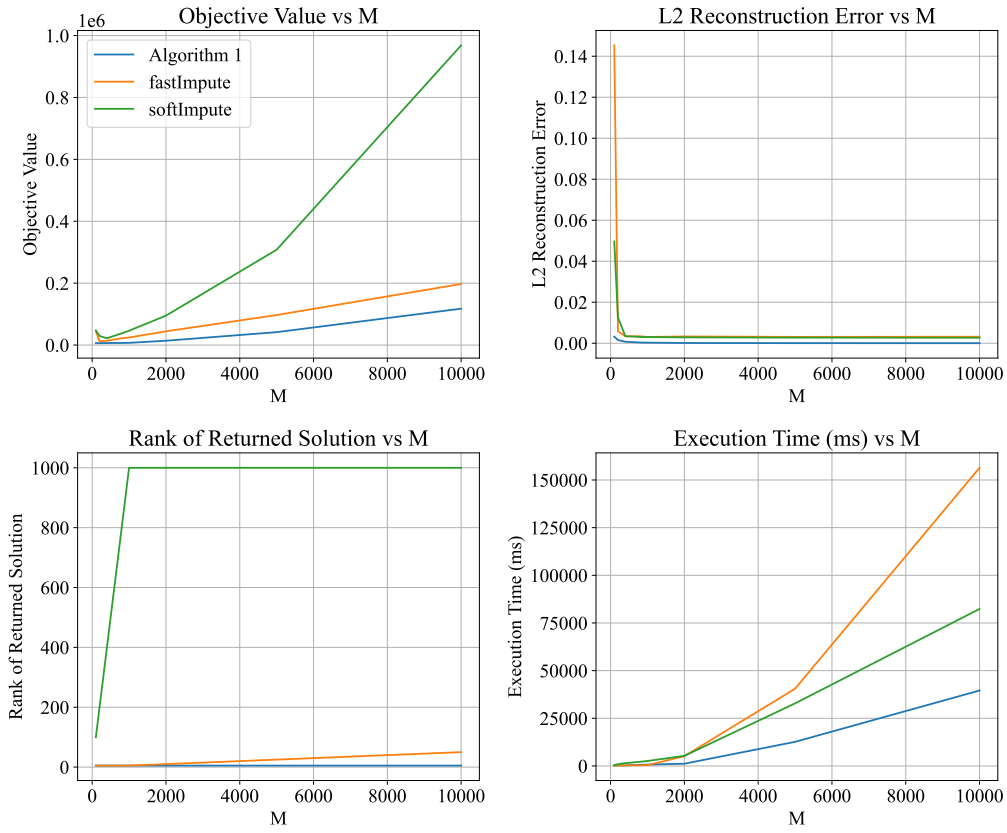

Figure 3: Objective value (top left), ℓ_2 reconstruction error (top right), fitted rank (bottom left) and execution time (bottom right) versus m with $n = 1000, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

objective value, reconstruction error and execution time for ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Tables 4, 5 and 6 of Appendix A. In Figure 4, we plot the average cumulative time spent solving subproblems (12), (14), (16), (18) during the execution of Algorithm 1 versus m. Our main findings from this set of experiments are a follows:

1. Here again, Algorithm 1 systematically produces higher quality solutions than ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD (see Table 4). On average, Algorithm 1 outputs a solution whose objective value is 62% lesser than the objective value achieved by the best performing alternative method (Fast-Impute). Here again, ScaledGD produces the weakest average objective value across these experiments. We observe that the objective value achieved by each method appears to increase super-linearly as the number of columns m increases.

2. In terms of ℓ_2 reconstruction error, Algorithm 1 again systematically produces solutions that are of higher quality than ScaledGD, Fast-Impute, Soft-Impute and

Iterative-SVD (see Table 5), often achieving an error that is an order of magnitude superior than the next best method. On average, Algorithm 1 outputs a solution whose ℓ_2 reconstruction error is 90% lesser than the reconstruction error achieved by the best performing alternative method (Soft-Impute in all but one parameter configuration).

3. The fitted rank of the solutions returned by Algorithm 1, ScaledGD and Fast-Impute always matched the specified target rank, but the solutions returned by Soft-Impute and Iterative-SVD were always of full rank despite the fact that these methods were provided with the target rank explicitly.

4. The runtime of Algorithm 1 exhibits the most favorable scaling behavior among the methods tested in these experiments. For instances with $m \geq 2000$, Table 6 shows that Algorithm 1 had the fastest runtime. For instances with $m < 2000$, ScaledGD had the fastest execution time but produced low quality solutions. The runtime of all methods tested grow super-linearly with m.

5. Figure 4 illustrates that the computation of the solution for (12) and (14) are the computational bottlenecks in the execution of Algorithm 1 in this set of experiments while the computation of the solution for (18) and (16) appear to be a constant function of m. This observation is consistent with the complexity analysis performed for each subproblem of Algorithm 1 in Section 5. Indeed, this analysis indicated that solve times for (18) and (16) are independent of m while the solve times for (12) and (14) scale linearly with m when the number of threads w satisfies $w < m$.

6.4 Sensitivity to Side Information Dimension

We present a comparison of Algorithm 1 with ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD as we vary the dimension of the side information d. In these experiments, we fixed $n = 1000, m = 100$ and $k = 5$ across all trials. We varied $d \in \{10, 50, 100, 150, 200, 250, 500, 1000\}$ and we performed 20 trials for each value of d.

We report the objective value, ℓ_2 reconstruction error, fitted rank and execution time for Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Figure 5. We additionally report the objective value, reconstruction error and execution time for ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Tables 7, 8 and 9 of Appendix A. In Figure 6, we plot the average cumulative time spent solving subproblems (12), (14), (16), (18) during the execution of Algorithm 1 versus d. Our main findings from this set of experiments are:

1. Just as in Sections 6.2 and 6.3, Algorithm 1 systematically produces higher quality solutions than ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD (see Table 7). On average, Algorithm 1 outputs a solution whose objective value is 85% lesser than the objective value achieved by the best performing alternative method (Fast-Impute). ScaledGD produces the weakest average objective value across these experiments. The objective value achieved by each method appears to increase linearly as the dimension d of the side information increases.

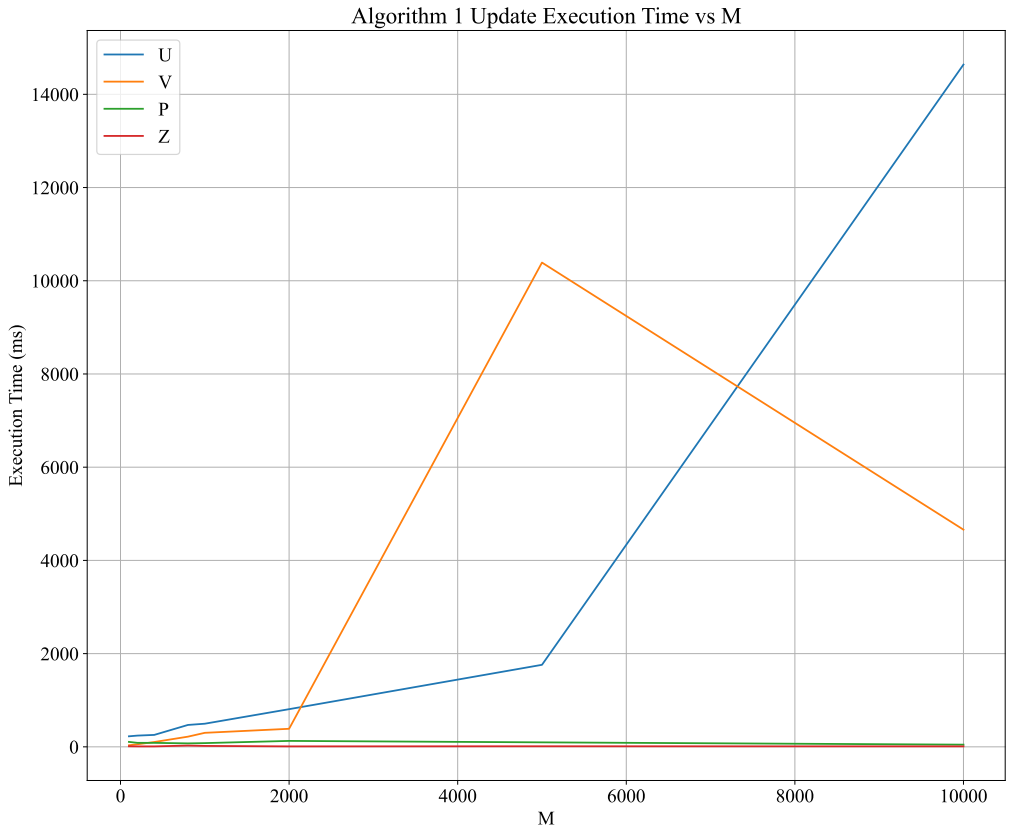

Figure 4: Cumulative time spent solving each subproblem of Algorithm 1 versus m with $n = 1000, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

2. In terms of ℓ_2 reconstruction error, just as in Sections 6.2 and 6.3 Algorithm 1 produces solutions that are of higher quality than ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD (see Table 8), often achieving an error that is an order of magnitude superior than the next best method. On average, Algorithm 1 outputs a solution whose ℓ_2 reconstruction error is 93% lesser than the reconstruction error achieved by the best performing alternative method (Soft-Impute). The performance of Algorithm 1 improves as d increases, consistent with the intuition that recovering the partially observed matrix $\boldsymbol{A}$ becomes easier as more side information becomes available.

3. The runtime of Algorithm 1 is competitive with that of the other methods. The runtime of Algorithm 1 is less than of Soft-Impute and Iterative-SVD but greater than that of Fast-Impute. Table 9 illustrates that ScaledGD was the fastest performing method, however its solutions were of the lowest quality. The runtime of Algorithm 1 and ScaledGD grows with d while Fast-Impute, Soft-Impute and iterate SVD are

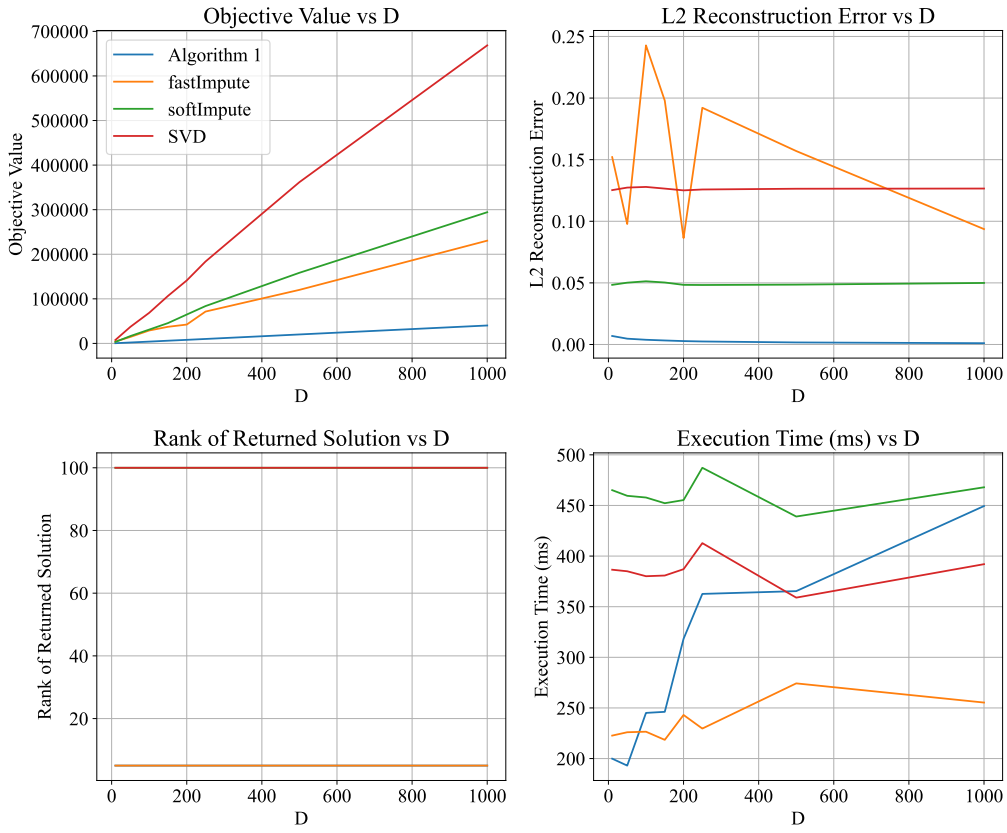

Figure 5: Objective value (top left), ℓ_2 reconstruction error (top right), fitted rank (bottom left) and execution time (bottom right) versus d with $n = 1000, m = 100$ and $k = 5$. Averaged over 20 trials for each parameter configuration.

constant with d which should be expected as these methods do not act on the side information matrix $\boldsymbol{Y}$.

4. Figure 6 illustrates that the computation of the solution for (16) is the computational bottleneck in the execution of Algorithm 1 in this set of experiments, followed next by the computation of the solution to (12). The solution times for (12), (14) and (18) appear constant as a function of d. This is consistent with the complexity analysis from Section 5 which found that the solve time for (16) is linear in d while the solve time for the 3 other subproblems are independent of d.

6.5 Sensitivity to Target Rank

We present a comparison of Algorithm 1 with ScaledGD, Fast-Impute, Soft-Impute and Iterative-SVD as we vary the rank of the underlying matrix k. In these experiments, we fixed

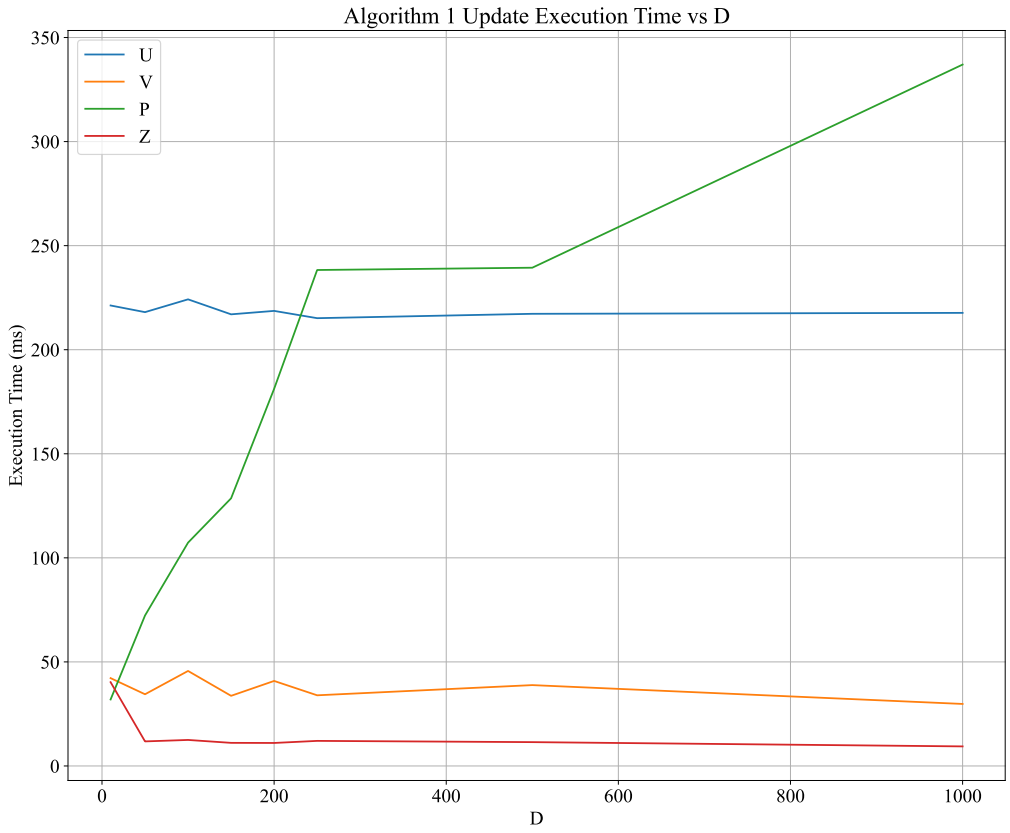

Figure 6: Cumulative time spent solving each subproblem of Algorithm 1 versus d with $n = 1000, m = 100$ and $k = 5$. Averaged over 20 trials for each parameter configuration.

$n = 1000, m = 100$ and $d = 150$ across all trials. We varied $k \in \{5, 10, 15, 20, 25, 30, 35, 40\}$ and we performed 20 trials for each value of d.

We report the objective value, ℓ_2 reconstruction error, fitted rank and execution time for Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Figure 7. We additionally report the objective value, reconstruction error and execution time for ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and Iterative-SVD in Tables 10, 11 and 12 of Appendix A. In Figure 8, we plot the average cumulative time spent solving subproblems (12), (14), (16), (18) during the execution of Algorithm 1 versus k. Our main findings from this set of experiments are as follows:

1. Unlike in Sections 6.2, 6.3 and 6.4, Algorithm 1 only produced higher quality solutions than all benchmark methods in 3 out of 8 of the tested parameter configurations where $k \leq 15$ (see Table 10). Fast-Impute was the best performing method in 3 configurations and Soft-Impute was best in the remaining 2 configurations. ScaledGD produces the weakest average objective value across these experiments.

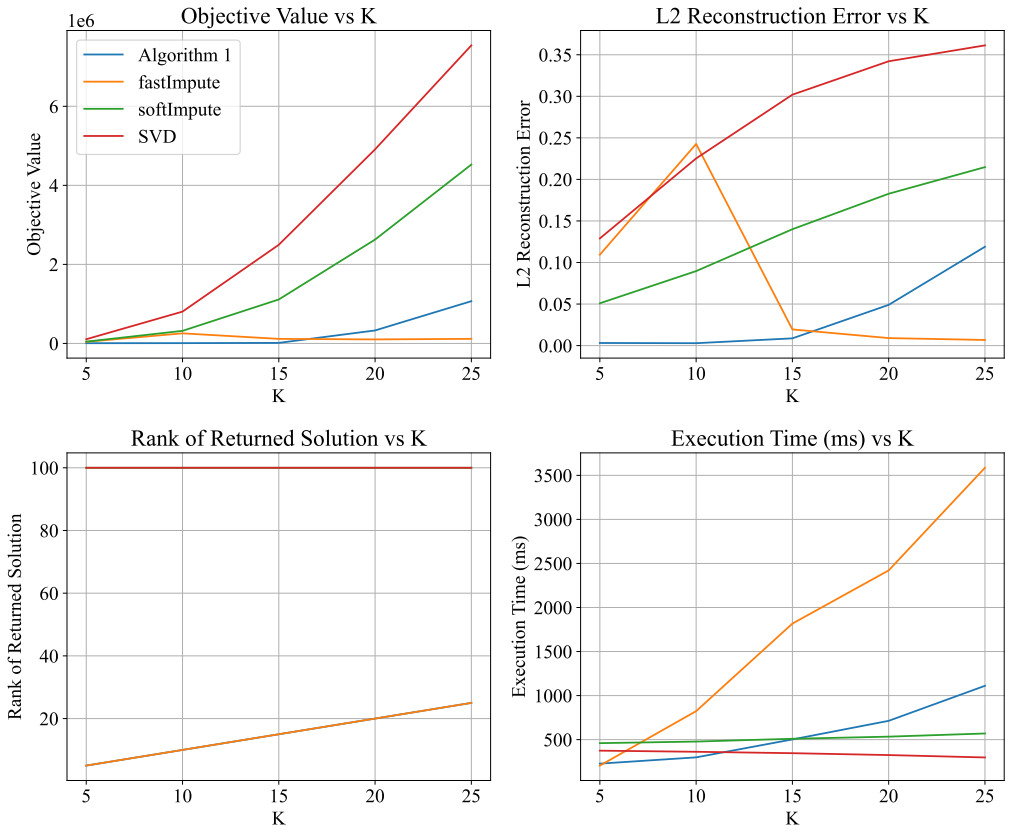

Figure 7: Objective value (top left), ℓ_2 reconstruction error (top right), fitted rank (bottom left) and execution time (bottom right) versus k with $n = 1000, m = 100$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

2. In terms of ℓ_2 reconstruction error, Algorithm 1 again produced higher quality solutions than all benchmark methods in 3 out of 8 of the tested parameter configurations where $k \leq 15$ (see Table 11). Fast-Impute produced solutions achieving the lowest error in the other 5 parameter configurations.

3. The runtime of Algorithm 1 is competitive with that of the other methods. Table 12 illustrates that ScaledGD was the fastest performing method, however its solutions were of the lowest quality. The runtime of Algorithm 1 is most competitive with Soft-Impute and Iterative-SVD for small values of k. Though Fast-Impute is the best performing method in terms of objective in 3 out of 8 configurations and the best in terms of ℓ_2 error in 5 out of 8 configurations, it takes on average 3 times as long as Algorithm 1 to execute.

4. Figure 8 illustrates that the computation of the solution for (12) is the computational bottleneck in the execution of Algorithm 1 in this set of experiments, followed next by the computation of the solution to (14) and (18).

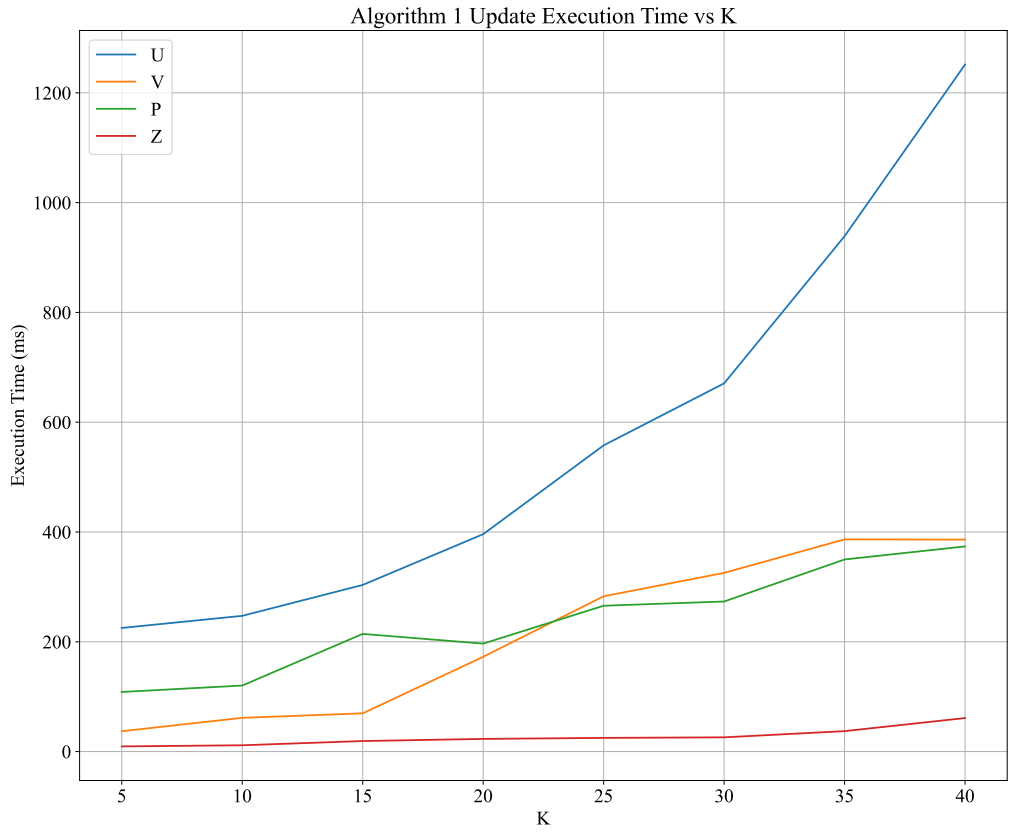

Figure 8: Cumulative time spent solving each subproblem of Algorithm 1 versus k with $n = 1000, m = 100$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

6.6 Summary of Findings

We now summarize our findings from our numerical experiments. In Sections 6.2-6.5, we see that across all experiments using synthetic data and target rank $k \leq 15$, Algorithm 1 produces solutions that achieve on average 79% lower objective value and 90.1% lower ℓ_2 reconstruction error than the solutions returned by the experiment-wise best performing benchmark method. In the regime where $k > 15$, we see in Section 6.5 that Fast-Impute outperforms Algorithm 1. We see that the execution time of Algorithm 1 is competitive with and often notably faster than the benchmark methods on synthetic data. Importantly, in the regime $k > 15$, although Fast-Impute returns higher quality solutions than Algorithm 1, the former has an execution time that is on average 3 times as long as our method. Our computational results are consistent with the complexity analysis performed in Section 5 for Problems (12), (14), (16) and (18). We observe that solution time for (16) becomes the bottleneck as the target rank k scales, otherwise the solution time for (12) is the bottleneck.

7. Conclusion

In this paper, we introduced Problem (1) which seeks to reconstruct a partially observed matrix that is predictive of fully observed side information. We illustrate that (1) has a natural interpretation as a robust optimization problem and can be reformulated as a mixed-projection optimization problem. We derive a semidefinite cone relaxation (9) to (1) and we present Algorithm 1, a mixed-projection alternating direction method of multipliers algorithm that obtains scalable, high quality solutions to (1). We rigorously benchmark the performance of Algorithm 1 on synthetic data against benchmark methods Fast-Impute, Soft-Impute, Iterative-SVD and ScaledGD. We find that across all experiments with $k \leq 15$, Algorithm 1 outputs solutions that achieve on average 79% lower objective value in (1) and 90.1% lower ℓ_2 reconstruction error than the solutions returned by the experiment-wise best performing benchmark method. For the 5 experiments with $k > 15$, Fast-Impute returns superior quality solutions than Algorithm 1, however the former takes on average 3 times as long as Algorithm 1 to execute. The runtime of Algorithm 1 is competitive with and often superior to that of the benchmark methods. Algorithm 1 is able to solve problems with $n = 10000$ rows and $m = 10000$ columns in less than a minute. Future work could expand the mixed-projection ADMM framework introduced in this work to incorporate positive semidefinite constraints and general linear constraints. Additionally, future work could empirically investigate the strength of the semidefinite relaxation (9) and could explore how to leverage this lower bound to certify globally optimal solutions.

References

Francis R Bach and Michael I Jordan. Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd international conference on Machine learning*, pages 33–40, 2005.

Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

Dimitris Bertsimas and Martin S Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931–942, 2018.

Dimitris Bertsimas and Dick den Hertog. *Robust and adaptive optimization*. Dynamic Ideas LLC, 2020.

Dimitris Bertsimas and Nicholas A. G. Johnson. Compressed sensing: A discrete optimization approach. *Machine Learning*, 2024. URL `https://doi.org/10.1007/s10994-024-06577-0`.

Dimitris Bertsimas and Michael Lingzhi Li. Fast exact matrix completion: A unified optimization framework for matrix completion. *Journal of Machine Learning Research*, 21 (231):1–43, 2020.

Dimitris Bertsimas, Jean Pauphilet, and Bart Van Parys. Rejoinder: Sparse regression: scalable algorithms and empirical performance. 2020.

Dimitris Bertsimas, Ryan Cory-Wright, and Jean Pauphilet. A unified approach to mixed-integer optimization problems with logical constraints. *SIAM Journal on Optimization*, 31(3):2340–2367, 2021.

Dimitris Bertsimas, Ryan Cory-Wright, and Jean Pauphilet. Mixed-projection conic optimization: A new paradigm for modeling rank constraints. *Operations Research*, 70(6): 3321–3344, 2022.

Dimitris Bertsimas, Ryan Cory-Wright, and Nicholas A. G. Johnson. Sparse plus low rank matrix decomposition: A discrete optimization approach. *Journal of Machine Learning Research*, 24(267):1–51, 2023a. URL `http://jmlr.org/papers/v24/21-1130.html`.

Dimitris Bertsimas, Ryan Cory-Wright, Sean Lo, and Jean Pauphilet. Optimal low-rank matrix completion: Semidefinite relaxations and eigenvector disjunctions. *arXiv preprint arXiv:2305.12292*, 2023b.

Dimitris Bertsimas, Ryan Cory-Wright, and Jean Pauphilet. A new perspective on low-rank optimization. *Mathematical Programming, articles in advance*, pages 1–46, 2023c.

Daniel Billsus, Michael J Pazzani, et al. Learning collaborative information filters. In *Icml*, volume 98, pages 46–54, 1998.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, USA, 2004.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. *Advances in neural information processing systems*, 28, 2015.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):301–337, 1995.

Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015.

Oktay Günlük and Jeff Linderoth. Perspective reformulation and applications. In *Mixed Integer Nonlinear Programming*, pages 61–89. Springer, 2012.

Ke Guo, Deren Han, David ZW Wang, and Tingting Wu. Convergence of admm for multiblock nonconvex separable optimization models. *Frontiers of Mathematics in China*, 12: 1139–1162, 2017.

Prateek Jain and Praneeth Netrapalli. Fast exact matrix completion with finite samples. In *Conference on Learning Theory*, pages 1007–1034. PMLR, 2015.

Hui Ji, Chaoqiang Liu, Zuowei Shen, and Yuhong Xu. Robust video denoising using low rank matrix completion. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1791–1798. IEEE, 2010.

Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29, 2016.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Qi Liu. Power network system identification and recovery based on the matrix completion. In *Journal of Physics: Conference Series*, volume 1237, page 032059. IOP Publishing, 2019.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80): 2287–2322, 2010. URL `http://jmlr.org/papers/v11/mazumder10a.html`.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate o (1/k2). In *Dokl akad nauk Sssr*, volume 269, page 543, 1983.

Luong Trung Nguyen, Junhan Kim, and Byonghyo Shim. Low-rank matrix completion: A contemporary survey. *IEEE Access*, 7:94215–94237, 2019.

Art B. Owen and Patrick O. Perry. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564 – 594, 2009.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Andy Ramlatchan, Mengyun Yang, Quan Liu, Min Li, Jianxin Wang, and Yaohang Li. A survey of matrix completion methods for recommendation systems. *Big Data Mining and Analytics*, 1(4):308–323, 2018.

Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, Michael Jones, Anna Klein, Lauren Milechin, Julia Mullen, Andrew Prout, Antonio Rosa, Charles Yee, and Peter Michaleas. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.

Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python, 2016. URL `https://github.com/iskandr/fancyimpute`.

Badrul Sarwar, George Karypis, Joseph Konstan, and John T Riedl. Application of dimensionality reduction in recommender system-a case study. 2000.

Anureet Saxena, Pierre Bonami, and Jon Lee. Convex relaxations of non-convex mixed integer quadratically constrained programs: extended formulations. *Mathematical programming*, 124(1):383–411, 2010.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Junxiang Wang and Liang Zhao. Nonconvex generalization of alternating direction method of multipliers for nonlinear equality constrained problems. *Results in Control and Optimization*, 2:100009, 2021.

Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78:29–63, 2019.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(7), 2009.

Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. *Advances in neural information processing systems*, 26, 2013.

Zheng Xu, Soham De, Mario Figueiredo, Christoph Studer, and Tom Goldstein. An empirical study of admm for nonconvex problems. *arXiv preprint arXiv:1612.03349*, 2016.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Appendix A. Supplemental Computational Results

Table 1: Comparison of the objective value of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus n with $m = 100, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	Objective				
N	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
100	249262.99	**655.46**	5530.48	19893.08	28677.73
200	306738.68	**1280.82**	9756.14	24251.93	44054.89
400	417643.27	**2483.49**	14321.46	28932.85	61112.62
800	421032.49	**4813.05**	31520.96	41179.38	93119.34
1000	522586.08	**6010.34**	33557.75	47701.68	107851.28
2000	563033.20	**11975.48**	83669.84	76566.52	167458.68
5000	1226489.68	**30060.61**	273747.04	172093.66	364065.64
10000	1973665.62	**60082.27**	521189.88	319915.98	642759.84

Table 2: Comparison of the reconstruction error of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus n with $m = 100, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	ℓ_2 Reconstruction Error				
N	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
100	100.22460	**0.01520**	0.07540	0.21330	0.30090
200	58.37210	**0.00695**	0.12800	0.11770	0.21050
400	33.92500	**0.00412**	0.08980	0.07390	0.15980
800	14.97890	**0.00328**	0.23990	0.05160	0.13010
1000	12.66500	**0.00312**	0.18600	0.04950	0.12620
2000	5.54420	**0.00304**	0.07990	0.04410	0.11670
5000	2.47260	**0.00282**	0.15040	0.03720	0.10550
10000	1.32070	**0.00267**	0.10920	0.03510	0.10240

Table 3: Comparison of the execution time of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus n with $m = 100, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	Execution Time (ms)				
N	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
100	**10.84**	53.47	99.95	141.42	115.26
200	**41.11**	73.84	121.05	187.26	163.11
400	**54.95**	113.89	152.16	262.95	246.42
800	**68.00**	184.11	195.05	389.63	334.16
1000	**44.11**	241.16	211.63	442.05	366.53
2000	**124.47**	340.84	311.32	759.58	575.79
5000	813.53	770.11	**484.58**	1795.58	1298.63
10000	18828.21	1730.00	**863.84**	3871.53	2713.26

Table 4: Comparison of the objective value of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus m with $n = 1000, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	Objective				
M	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
100	530097.31	**6014.93**	44337.08	47334.20	103403.04
200	2483913.82	**6131.52**	12159.08	29560.77	114448.43
400	14226534.31	**6361.90**	13875.31	21942.31	90652.40
800	99356634.18	**6800.63**	22152.43	37924.99	87895.33
1000	105451997.06	**7126.60**	24435.40	46327.69	128499.51
2000	591164404.77	**13964.62**	44333.51	95044.30	815807.16
5000	4002087935.12	**41679.23**	96985.27	308044.93	11294104.83
10000	9826251365.01	**117558.76**	197362.37	968255.00	60913874.17

Table 5: Comparison of the reconstruction error of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus m with $n = 1000, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	ℓ_2 Reconstruction Error				
M	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
100	13.68740	**0.00322**	0.14530	0.04960	0.12560
200	40.58900	**0.00154**	0.00590	0.01260	0.06640
400	127.71450	**0.00075**	0.00340	0.00340	0.02240
800	508.24550	**0.00036**	0.00340	0.00310	0.00460
1000	443.26620	**0.00029**	0.00310	0.00300	0.00350
2000	1292.61610	**0.00012**	0.00330	0.00290	0.00300
5000	3658.18810	**0.00004**	0.00310	0.00270	0.00540
10000	4559.27360	**0.00002**	0.00320	0.00270	0.00650

Table 6: Comparison of the execution time of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus m with $n = 1000, k = 5$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	Execution Time (ms)				
M	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
100	**44.16**	222.21	193.16	444.89	365.74
200	**54.32**	237.37	223.32	953.32	760.11
400	**80.95**	311.32	315.47	1466.32	1511.79
800	**121.37**	637.53	360.53	2198.21	2564.00
1000	**154.89**	728.58	434.47	2611.58	3009.21
2000	4652.11	**1181.47**	5127.37	5308.16	6062.89
5000	28587.11	**12645.16**	40526.16	32824.79	35015.21
10000	108255.05	**39569.37**	156399.42	82361.37	86762.84

Table 7: Comparison of the objective value of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus d with $n = 1000, m = 100$ and $k = 5$. Averaged over 20 trials for each parameter configuration.

	Objective				
D	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
10	11691.78	**475.17**	4222.64	3367.41	7710.45
50	96771.27	**2067.20**	14565.83	16511.97	37203.79
100	229740.41	**4033.46**	28967.06	31000.81	68634.31
150	532018.26	**6057.23**	37244.64	45581.92	106504.32
200	734648.30	**7994.51**	42289.45	64771.47	141252.05
250	1195065.81	**9984.20**	71433.91	83752.90	183720.00
500	4165782.61	**20094.24**	120114.14	158458.64	361740.57
1000	13578263.54	**40191.29**	230467.93	294273.13	668550.27

Table 8: Comparison of the reconstruction error of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus d with $n = 1000, m = 100$ and $k = 5$. Averaged over 20 trials for each parameter configuration.

	ℓ_2 Reconstruction Error				
D	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
10	0.60780	**0.00690**	0.15210	0.04840	0.12530
50	1.41600	**0.00471**	0.09780	0.05020	0.12730
100	5.03590	**0.00382**	0.24280	0.05130	0.12790
150	14.00330	**0.00326**	0.19790	0.05030	0.12650
200	22.26870	**0.00276**	0.08640	0.04840	0.12510
250	39.41630	**0.00245**	0.19210	0.04830	0.12580
500	177.68800	**0.00165**	0.15700	0.04860	0.12640
1000	679.11770	**0.00104**	0.09360	0.04990	0.12660

Table 9: Comparison of the execution time of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus d with $n = 1000, m = 100$ and $k = 5$. Averaged over 20 trials for each parameter configuration.

	Execution Time (ms)				
D	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
10	**84.00**	199.89	222.68	465.11	386.53
50	**80.79**	193.05	226.00	459.53	385.00
100	**108.63**	245.11	226.53	457.84	380.11
150	**113.47**	246.16	218.47	452.21	380.79
200	**117.32**	318.26	243.05	455.32	387.05
250	**152.79**	362.63	229.63	487.21	412.79
500	**176.21**	365.37	274.26	439.05	358.95
1000	**138.74**	449.42	255.32	467.95	392.00

Table 10: Comparison of the objective value of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus k with $n = 1000, m = 100$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	Objective				
K	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
5	514330.09	**6021.76**	41376.50	45858.20	106390.58
10	1892278.99	**7393.65**	255228.60	318398.98	805396.86
15	5213393.44	**14104.62**	115383.43	1112396.97	2495972.46
20	10196279.89	328671.00	**101812.52**	2628073.04	4910386.51
25	16816442.74	1069103.04	**116005.02**	4526388.13	7541300.54
30	-	34567679.83	10695127.17	**6864577.33**	10634436.81
35	39536651.09	187701091.79	**144464.25**	9424715.13	14192827.60
40	-	723504611.03	191276652.97	**12529277.05**	18290215.22

Table 11: Comparison of the reconstruction error of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus k with $n = 1000, m = 100$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	ℓ_2 Reconstruction Error				
K	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
5	12.95500	**0.00314**	0.10940	0.05090	0.12900
10	5.41720	**0.00288**	0.24270	0.08970	0.22520
15	3.45210	**0.00871**	0.01940	0.14010	0.30200
20	2.34590	0.04900	**0.00903**	0.18270	0.34220
25	1.73610	0.11890	**0.00678**	0.21480	0.36130
30	-	0.20030	**0.00562**	0.23810	0.37240
35	1.16780	0.17330	**0.00488**	0.25260	0.37740
40	-	0.22320	**0.00452**	0.26550	0.38020

Table 12: Comparison of the execution time of ScaledGD, Algorithm 1, Fast-Impute, Soft-Impute and SVD versus k with $n = 1000, m = 100$ and $d = 150$. Averaged over 20 trials for each parameter configuration.

	Execution Time (ms)				
K	ScaledGD	Algorithm 1	Fast-Impute	Soft-Impute	SVD
5	**55.79**	227.47	205.16	460.11	374.79
10	**80.79**	298.63	823.21	477.95	361.95
15	**107.89**	502.21	1817.16	509.68	346.05
20	**95.53**	713.42	2420.89	533.79	324.79
25	**111.53**	1110.89	3586.05	569.68	297.58
30	-	1353.95	4435.63	591.21	**280.37**
35	**107.21**	1822.16	6212.63	640.05	271.00
40	-	2281.95	8168.68	645.11	**263.37**