Noisy Matrix Decomposition via Convex Relaxation: Optimal Rates in High Dimensions

Alekh Agarwal   ALEKH@EECS.BERKELEY.EDU
Sahand Negahban   SAHAND_N@EECS.BERKELEY.EDU
Department of EECS, University of California, Berkeley, CA 94720, USA

Martin J. Wainwright   WAINWRIG@STAT.BERKELEY.EDU
Departments of Statistics and EECS, University of California, Berkeley, CA 94720, USA

Abstract

We analyze a class of estimators based on a convex relaxation for solving high-dimensional matrix decomposition problems. The observations are the noisy realizations of the sum of an (approximately) low rank matrix $\Theta^\star$ with a second matrix $\Gamma^\star$ endowed with a complementary form of low-dimensional structure. We derive a general theorem that gives upper bounds on the Frobenius norm error for an estimate of the pair $(\Theta^\star, \Gamma^\star)$ obtained by solving a convex optimization problem. We then specialize our general result to two cases that have been studied in the context of robust PCA: low rank plus sparse structure, and low rank plus a column sparse structure. Our theory yields Frobenius norm error bounds for both deterministic and stochastic noise matrices, and in the latter case, they are minimax optimal. The sharpness of our theoretical predictions is also confirmed in numerical simulations.

1. Introduction

In this paper, we study a class of high-dimensional *matrix decomposition* problems. Suppose that we observe a matrix $Y \in \mathbb{R}^{d_1 \times d_2}$ that is (approximately) equal to the sum of two unknown matrices: how to recover good estimates of the pair? Of course, this problem is ill-posed in general, so that it is necessary to impose some kind of low-dimensional structure on the matrix components, one example being rank constraints. The framework of this paper supposes that one matrix component (denoted $\Theta^\star$) is low-rank, ei-

ther exactly or in an approximate sense, and allows for general forms of low-dimensional structure for the second component $\Gamma^\star$. Two particular cases of structure for $\Gamma^\star$ that have been considered in past work are elementwise sparsity (Chandrasekaran et al., 2009; 2010; Candes et al., 2009; Hsu et al., 2010) and column-wise sparsity (McCoy & Tropp, 2010; Xu et al., 2010).

Problems of matrix decomposition are motivated by a variety of applications. Many classical methods for dimensionality reduction, among them principal components analysis (PCA), are based on estimating a low-rank matrix from data. It is natural to ask whether or not such estimation procedures are robust to errors. Particularly harmful are gross errors that might affect only a few observations, but in an uncontrolled and potentially even adversarial fashion. This leads to different forms of robust PCA, which can be formulated in terms of matrix decomposition (Chandrasekaran et al., 2009; Candes et al., 2009; Hsu et al., 2010; Xu et al., 2010), using the matrix $\Gamma^\star$ to model the adversarial errors. Other applications include controlling sensor failures in video and image processing (Candes et al., 2009), performing figure-background separation in video processing, and dealing with malicious users in recommendation systems (Xu et al., 2010). Related decompositions arise in Gaussian covariance selection with hidden variables (Chandrasekaran et al., 2010).

More concretely, this paper focuses on matrix decompositions of the form

$$Y = \Theta^\star + \Gamma^\star + W, \qquad (1)$$

where W is some type of observation noise that is potentially dense, and can either be deterministic or stochastic. The matrix $\Theta^\star$ is assumed to be either exactly low-rank, or well-approximated by a low-rank matrix, whereas the matrix $\Gamma^\star$ is assumed to have a complementary type of low-dimensional structure, such as sparsity. Our goal is to recover accurate es-

timates of the decomposition $(\Theta^\star, \Gamma^\star)$ based on the noisy observations Y. In this paper, we analyze simple estimators based on convex relaxations involving the nuclear norm, and a second general norm $\mathcal{R}$.

Some interesting special cases of the model (1) have been examined in some recent work, almost exclusively in the noiseless setting $(W = 0)$. Chandrasekaran et al. (2009) studied the case when $\Gamma^\star$ is assumed to be sparse, with a relatively small number s of non-zero entries. In the noiseless setting, they gave sufficient conditions for exact recovery for an adversarial sparsity model, meaning the non-zero positions of $\Gamma^\star$ can be arbitrary. Subsequent work by Candes et al. (2009) studied the same model but under a random sparsity model, in which the non-zero positions are chosen uniformly at random. Most closely related is the work of Hsu et al. (2010), who also analyze the noisy case with elementwise sparsity, but require the rank of $\Gamma^\star$ to be constrained by $rs \leq d_1 d_2/(\log d_1 \log d_2)$; this scaling is not minimax-optimal, in contrast to the results presented here. In very recent work, Xu et al. (2010) proposed and analyzed a different column-sparse model, in which the matrix $\Gamma^\star$ has a relatively small number $s \ll d_2$ of non-zero columns.

In this paper, we study a general class of matrix decomposition problems that include these models as special cases. Our main contribution is to provide a general result (Theorem 1) on approximate recovery of the unknown decomposition from noisy observations, valid for a fairly general class of structural constraints on $\Gamma^\star$ (explained in next section). The upper bound in Theorem 1 consists of multiple terms, each of which has a natural interpretation in terms of the estimation and approximation errors associated with the subproblems of recovering $\Theta^\star$ and $\Gamma^\star$. We then specialize this general result to the case of elementwise or column-wise sparsity models for $\Gamma^\star$, thereby obtaining recovery guarantees for matrices $\Theta^\star$ that may be either exactly or approximately low-rank, as well as matrices $\Gamma^\star$ that may be either exactly or approximately sparse. In addition, our results hold for general noisy observations $(W \neq 0)$. To the best of our knowledge, these are the first results that apply to this broad class of models. Moreover, the rates obtained by our analysis cannot be improved in general. Indeed, when the noise matrix W is stochastic, our results can be shown to be minimax-optimal up to constant factors.

An interesting feature of our analysis is that, in contrast to the past work just discussed, we do *not* impose incoherence conditions on the singular vectors of $\Theta^\star$; rather, we control the interaction with a milder con-

dition involving the dual norm of the regularizer. In the special case of elementwise sparsity, this dual norm enforces an upper bound on the "spikiness" of the low-rank component. As we show, this type of milder constraint is both necessary and sufficient for the approximate recovery that is of interest in the noisy setting.

The remainder of the paper is organized as follows. In Section 2, we set up the problem in a precise way, and describe the estimators. Section 3 is devoted to the statement of our main result, as well as its various corollaries for special cases of the matrix decomposition problem. Section 4 provides proofs of the main corollaries. In Section 5, we provide numerical simulations that illustrate the sharpness of our theoretical predictions. Proofs of the main theorem and other technical results can be found in the full-length version of this paper (Agarwal et al., 2011).

2. Convex relaxations and matrix decomposition

In this paper, we consider a family of regularizers formed by a combination of the *nuclear norm*

$$\|\Theta\|_{\mathrm{N}} := \sum_{j=1}^{\min\{d_1,d_2\}} \sigma_j(\Theta), \qquad (2)$$

or the sum of singular values of Θ, which acts as a convex surrogate to a rank constraint for $\Theta^\star$ (see e.g. Recht et al. (2010) and references therein), with a *norm-based regularizer* $\mathcal{R} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_+$ used to constrain the structure of $\Gamma^\star$. We provide a general theorem applicable to a class of regularizers $\mathcal{R}$ that satisfy a certain decomposability property (Negahban et al., 2009), and then consider in detail a few particular choices of $\mathcal{R}$ that have been studied in past work, including the elementwise ℓ_1-norm, and the columnwise $(1, 2)$-norm (see Examples 1 and 2 below).

In the presence of noise $(W \neq 0)$, it is natural to consider the family of estimators

$$\min_{(\Theta,\Gamma)} \left\{ \frac{1}{2} \|Y - (\Theta + \Gamma)\|_{\mathrm{F}}^2 + \lambda_d \|\Theta\|_{\mathrm{N}} + \mu_d \mathcal{R}(\Gamma) \right\},$$

where (λ_d, μ_d) are non-negative regularization parameters, to be chosen by the user, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Our theory also provides choices of these parameters that guarantee good properties of the associated estimator. Although the estimator is reasonable, it turns out that an additional constraint yields an equally simple estimator that has attractive properties, both in theory and in practice.

In order to understand the need for an additional constraint, it should be noted that, without further con-

straints, the model (1) is unidentifiable even in the noiseless setting ($W = 0$). As has been discussed in past work, no method can recover the components $(\Theta^\star, \Gamma^\star)$ unless the low-rank component is "incoherent" with the matrix $\Gamma^\star$. For instance, supposing for the moment that $\Gamma^\star$ is a sparse matrix, consider a rank one matrix with $\Theta^\star_{11} \neq 0$, and zeros in all other positions. In this case, it is clearly impossible to disentangle $\Theta^\star$ from a sparse matrix. Past work on both matrix completion and decomposition has ruled out these types of troublesome cases via conditions on the singular vectors of the low-rank component $\Theta^\star$, and used them to derive sufficient conditions for exact recovery in the noiseless setting. In this paper, we impose a milder condition with the goal of performing approximate recovery. It should be noted that in the more realistic setting of noisy observations and/or matrices that are not exactly low-rank, such approximate recovery is the best that can be expected, and indeed, we also show that our rates are minimax-optimal, meaning that no algorithm can do substantially better.

For a given regularizer $\mathcal{R}$, we define the quantity $\kappa_d(\mathcal{R}) := \sup_{V \neq 0} \|V\|_F / \mathcal{R}(V)$, which measures the relation between the regularizer and the Frobenius norm. Moreover, we define the associated dual norm $\mathcal{R}^*(U) := \sup_{\mathcal{R}(V) \leq 1} \langle\!\langle V, U \rangle\!\rangle$, where $\langle\!\langle V, U \rangle\!\rangle := \text{trace}(V^T U)$ is the trace inner product on the space $\mathbb{R}^{d_1 \times d_2}$. Our estimators are based on constraining the interaction between the low-rank component $\Theta^\star$ and $\Gamma^\star$ via the quantity

$$\varphi_{\mathcal{R}}(\Theta) := \kappa_d(\mathcal{R}^*)\, \mathcal{R}^*(\Theta). \qquad (3)$$

With these definitions, the analysis of this paper is based on the family of estimators

$$\min_{(\Theta, \Gamma)} \left\{ \frac{1}{2} \|Y - (\Theta + \Gamma)\|_F^2 + \lambda_d \|\Theta\|_N + \mu_d \mathcal{R}(\Gamma) \right\}, \qquad (4)$$

subject to $\varphi_{\mathcal{R}}(\Theta) \leq \alpha$ for some fixed parameter α. Let us consider some examples to provide intuition.

Example 1 (Sparsity and elementwise ℓ_1-norm)**.** Suppose that $\Gamma^\star$ is assumed to be sparse, with $s \ll d_1 d_2$ non-zero entries. In this case, the sum $\Theta^\star + \Gamma^\star$ corresponds to the sum of a low rank matrix with a sparse matrix. This type of decomposition is motivated in various applications, among them robust forms of PCA (Candes et al., 2009) and Gaussian graph selection with hidden variables (Chandrasekaran et al., 2010). Since $\Gamma^\star$ is sparse, an appropriate choice of

regularizer is the elementwise ℓ_1-norm

$$\mathcal{R}(\Gamma) = \|\Gamma\|_1 := \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} |\Gamma_{jk}|. \qquad (5)$$

It is straightforward to verify that

$$\mathcal{R}^*(U) = \|U\|_\infty := \max_{j=1,\ldots,d_1} \max_{k=1,\ldots,d_2} |U_{jk}|, \qquad (6)$$

and moreover, that $\kappa_d(\mathcal{R}^*) = \sqrt{d_1 d_2}$. Consequently, in this specific case, the general convex program (4) takes the form

$$\min_{(\Theta, \Gamma)} \left\{ \frac{1}{2} \|Y - (\Theta + \Gamma)\|_F^2 + \lambda_d \|\Theta\|_N + \mu_d \|\Gamma\|_1 \right\}$$

$$\text{such that } \|\Theta\|_\infty \leq \frac{\alpha}{\sqrt{d_1 d_2}}. \qquad (7)$$

The constraint involving $\|\Theta\|_\infty$ serves to control the "spikiness" of the low rank component, with larger settings of α allowing for more spiky matrices; it arises in low-rank matrix completion problems (Negahban & Wainwright, 2010), as well as in the recent work of Hsu et al. (2010). More concretely, if we consider matrices with $\|\Theta\|_F \approx 1$, then setting $\alpha \approx 1$ allows only for matrices for which $|\Theta_{jk}| \approx 1/\sqrt{d_1 d_2}$ in all entries. If we want to permit the maximally spiky matrix with all its mass in a single position, then the parameter α must be of the order $\sqrt{d_1 d_2}$. In practice, we are interested in settings of α in between these two extremes.

♣

Example 2 (Column-sparsity and block columnwise regularization)**.** Motivated by robust PCA, Xu et al. (2010) have analyzed models in which $\Gamma^\star$ has a relatively small number $s \ll d_2$ of non-zero columns. In this case, it is natural to impose the $(1, 2)$-norm regularizer

$$\mathcal{R}(\Gamma) = \|\Gamma\|_{1,2} := \sum_{k=1}^{d_2} \|\Gamma_k\|_2, \qquad (8)$$

where Γ_k is the kth column of Γ. For this choice, it can be verified that

$$\mathcal{R}^*(U) = \|U\|_{\infty,2} := \max_{k=1,2,\ldots,d_2} \|U_k\|_2, \qquad (9)$$

where U_k denotes the kth column of U. Also, $\kappa_d(\mathcal{R}^*) = \sqrt{d_2}$. Consequently, in this specific case, the convex program (4) takes the form

$$\min_{(\Theta, \Gamma)} \left\{ \frac{1}{2} \|Y - (\Theta + \Gamma)\|_F^2 + \lambda_d \|\Theta\|_N + \mu_d \|\Gamma\|_{1,2} \right\}$$

$$\text{such that } \|\Theta\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}}. \qquad (10)$$

Again the constraint on $\|\Theta\|_{\infty,2}$ serves to limit the "spikiness" of the low rank component, where in this case, spikiness is measured columnwise. ♣

3. Main results and their consequences

In this section, we state our main results, and illustrate some of their consequences.

3.1. Decomposable regularizers

Our results apply to the family of convex programs (4) whenever the regularizer $\mathcal{R}$ is decomposable (Negahban et al., 2009). The notion of decomposability is defined in terms of a pair of subspaces, which (in general) need not be orthogonal complements. Here we consider a special case of decomposability:

Definition 1. Given a subspace $\mathcal{M} \subseteq \mathbb{R}^{d_1 \times d_2}$ and its orthogonal complement $\mathcal{M}^\perp$, a norm-based regularizer $\mathcal{R}$ is *decomposable with respect* $(\mathcal{M}, \mathcal{M}^\perp)$ if for all $U \in \mathcal{M}$, and $V \in \mathcal{M}^\perp$.

$$\mathcal{R}(U + V) = \mathcal{R}(U) + \mathcal{R}(V) \qquad (11)$$

To provide some intuition, the subspace $\mathcal{M}$ should be thought of as the nominal *model subspace*; in our results, it will be chosen such that the matrix $\Gamma^\star$ lies within or close to $\mathcal{M}$. The orthogonal complement $\mathcal{M}^\perp$ represents deviations away from the model subspace, and the equality (11) guarantees that such deviations are penalized as much as possible.

As discussed at more length by Negahban et al. (2009), a large class of norms are decomposable with respect to interesting[1] subspace pairs. Of particular relevance to us is the decomposability of the elementwise ℓ_1-norm $\|\Gamma\|_1$, and the columnwise $(1,2)$-norm $\|\Gamma\|_{1,2}$ discussed in Examples 1 and 2 respectively. Beginning with the elementwise ℓ_1-norm, given an arbitrary subset $S \subseteq \{1, 2, \ldots, d_1\} \times \{1, 2, \ldots, d_2\}$ of matrix indices consider the subspace pair

$$\mathcal{M}(S) := \left\{ U \in \mathbb{R}^{d_1 \times d_2} \mid U_{jk} = 0 \; \forall (j,k) \notin S \right\} \quad (12)$$

and its orthogonal complement $\mathcal{M}^\perp(S)$. It is easy to see that for any $U \in \mathcal{M}(S)$ and $V \in \mathcal{M}^\perp(S)$, we have $\|U + V\|_1 = \|U\|_1 + \|V\|_1$, showing that the elementwise ℓ_1-norm is decomposable with respect to the pair $(\mathcal{M}(S), \mathcal{M}^\perp(S))$. Similarly, it is straightforward to verify that the columnwise $(1,2)$-norm is decomposable with respect to subspace pairs defined analogously with $S \subseteq \{1, 2, \ldots, d_2\}$ a subset of column indices.

For any decomposable regularizer and non-trivial subspace $\mathcal{M}$, we define the compatibility constant

$$\Psi(\mathcal{M}, \mathcal{R}) := \sup_{U \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(U)}{\|U\|_F}. \qquad (13)$$

---

[1] Note that any norm is (trivially) decomposable with respect to the pair $(\mathcal{M}, \mathcal{M}^\perp) = (\mathbb{R}^{d_1 \times d_2}, \{0\})$.

For example, for the ℓ_1-norm and the set $\mathcal{M}(S)$ previously defined, an elementary calculation yields $\Psi(\mathcal{M}(S); \| \cdot \|_1) = \sqrt{|S|}$.

3.2. A general result

We begin by stating a result for a general decomposable regularizer $\mathcal{R}$, and a general noise matrix W. In later subsections, we specialize this result to particular choices of regularizers, and then to stochastic noise matrices. We denote the operator norm of a matrix, or equivalently its largest singular value, by $\|\cdot\|_{\mathrm{op}}$.

Theorem 1. *Given observations Y from the model (1), suppose that we solve the convex program (4) with regularization parameters (λ_d, μ_d) such that*

$$\lambda_d \geq 4\|W\|_{\mathrm{op}}, \; and \; \mu_d \geq 4\mathcal{R}^*(W) + \frac{4\,\alpha}{\kappa_d}. \qquad (14)$$

Then there is a universal constant c_1 such that for any matrix pair $(\Theta^\star, \Gamma^\star)$ with $\varphi_\mathcal{R}(\Theta^\star) \leq \alpha$, and for all integers $r = 1, 2, \ldots, \min\{d_1, d_2\}$, and any $\mathcal{R}$-decomposable pair $(\mathcal{M}, \mathcal{M}^\perp)$,

$$\underbrace{\|\widehat{\Theta} - \Theta^\star\|_F^2 + \|\widehat{\Gamma} - \Gamma^\star\|_F^2}_{e^2(\widehat{\Theta};\widehat{\Gamma})} \leq c_1 \mathcal{K}_{\Theta^\star}^\star + c_1 \mathcal{K}_\Gamma^\star, \qquad (15)$$

where

$$\mathcal{K}_\Theta^\star := \lambda_d^2 \left\{ r + \frac{1}{\lambda_d} \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^\star) \right\}, \quad and$$

$$\mathcal{K}_\Gamma^\star := c_1 \mu_d^2 \left\{ \Psi^2(\mathcal{M}; \mathcal{R}) + \frac{1}{\mu_d} \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\Gamma^\star)) \right\}.$$

Remarks: Note that the rate (15) is defined by a sum of two terms: the terms $\mathcal{K}_\Theta^\star$ and $\mathcal{K}_\Gamma^\star$ correspond (respectively) to the complexities associated with the sub-problems of recovering $\Theta^\star$ and $\Gamma^\star$. Each term is further divided into a sum of two sub-terms, which have an interpretation as the estimation error and the approximation error. Considering for instance the term $\mathcal{K}_\Theta^\star$, as will be clarified in the sequel, the term $\lambda_d^2 r$ corresponds to the *estimation error* associated with a rank r matrix, whereas the term $\lambda_d \sum_{j=r+1}^{\min\{d_1,d_2\}} \sigma_j(\Theta^\star)$ corresponds to the *approximation error* associated with representing $\Theta^\star$ (which might be full rank) by a matrix of rank r. A similar interpretation applies to the two components of $\mathcal{K}_\Gamma^\star$.

Since the inequality (15) corresponds to a family of upper bounds indexed by r and the subspace $\mathcal{M}$, these quantities can be chosen adaptively, depending on the structure of the matrices $(\Theta^\star, \Gamma^\star)$, so as to obtain the tightest possible upper bound. In the simplest case,

the matrix $\Theta^\star$ is exactly low rank (say rank r), and $\Gamma^\star$ lies within a $\mathcal{R}$-decomposable subspace $\mathcal{M}$. In this case, the approximation errors vanish, and Theorem 1 guarantees that the squared Frobenius error is at most

$$e^2(\widehat{\Theta}; \widehat{\Gamma}) \le c_1 \lambda_d^2 r + c_1 \mu_d^2 \Psi^2(\mathcal{M}; \mathcal{R}). \qquad (16)$$

3.3. Results for ℓ_1-norm regularization

Theorem 1 holds for any regularizer that is decomposable with respect to some subspace pair. As previously noted, an important example of a decomposable regularizer is the elementwise ℓ_1-norm, which is decomposable with respect to subspaces of the form (12). If we make the additional assumption that $\Theta^\star$ has rank r and $\Gamma^\star$ is exactly sparse, with at most $s \ll d_1 d_2$ entries, then the simplified form (16) implies immediately that

$$\|\widehat{\Theta} - \Theta^\star\|_F^2 + \|\widehat{\Gamma} - \Gamma^\star\|_F^2 \le c_1 \lambda_d^2 r + c_1 \mu_d^2 s,$$

where we have used the fact that $\Psi^2(\mathcal{M}(S); \|\cdot\|_1) = s$ for the model subspace defined by a subset S of cardinality s.

Further specializing to the case of exact observations $(W = 0)$, yields a form of approximate recovery—namely

$$\|\widehat{\Theta} - \Theta^\star\|_F^2 + \|\widehat{\Gamma} - \Gamma^\star\|_F^2 \lesssim \alpha^2 \frac{s}{d_1 d_2}. \qquad (17)$$

This guarantee is weaker than the exact recovery results obtained in other previous work; however, this past work imposed incoherence requirements on the singular vectors of the low-rank component $\Theta^\star$ that are more restrictive than the conditions of Theorem 1. It is important to note that *without* imposing these incoherence conditions, exact recovery is impossible, and moreover, the approximate recovery guarantee (17) cannot be improved. Indeed, consider the matrix $\Theta^\star = \frac{\alpha}{\sqrt{d_1 d_2}} \vec{1} f^T$, where the vector $f \in \mathbb{R}^{d_2}$ has $\frac{s}{d_1}$ entries equal to one, and the remainder zero. By construction, this matrix is rank one, satisfies $\|\Theta^\star\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}$, and $\|\Theta^\star\|_F^2 = \alpha^2 \frac{s}{d_1 d_2}$. In addition, it has at most s non-zero entries, meaning that under our model, an "adversary" could set $\Gamma^\star = -\Theta^\star$, so that we observe the all-zeroes matrix $Y = \Theta^\star + \Gamma^\star = 0$. Consequently, under the conditions of Theorem 1, no method can estimate to greater accuracy than $\frac{\alpha^2 s}{d_1 d_2}$, even in the noiseless setting $(W = 0)$.

Our discussion thus far has applied to general matrices W. More concrete results can be obtained by assuming that W is stochastic.

Corollary 1. *Suppose $\Theta^\star$ has rank at most r with $\|\Theta^\star\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}$, and $\Gamma^\star$ has at most s non-zero entries. If the noise matrix W has i.i.d. $N(0, \nu^2/(d_1 d_2))$*

entries, and we solve the convex program (7) with $\lambda_d = \frac{8\nu}{\sqrt{d_1}} + \frac{8\nu}{\sqrt{d_2}}$, and $\mu_d = 16\nu \sqrt{\frac{\log(d_1 d_2)}{d_1 d_2}} + \frac{4\alpha}{\sqrt{d_1 d_2}}$, then with probability greater than $1 - \exp\left(-2\log(d_1 d_2)\right)$, the error $e^2(\widehat{\Theta}, \widehat{\Gamma})$ of any solution is at most

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) \le c_1 \nu^2 \left(\frac{r}{d_1} + \frac{r}{d_2} + \frac{s \log(d_1 d_2)}{d_1 d_2}\right) + c_1 \frac{\alpha^2 s}{d_1 d_2}.$$

In this case the settings of λ_d, μ_d are based on upper bounding $\|W\|_{\mathrm{op}}$ and $\|W\|_\infty$. With a slightly modified argument, this bound can be sharpened by reducing the logarithmic term to $\log(\frac{d_1 d_2}{s})$. This bound is minimax-optimal, meaning that no estimator (regardless of its computational complexity) can achieve much better estimates for the matrix classes and noise model given here, which we further discuss in section 3.5.

3.4. Results for $\|\cdot\|_{1,2}$ regularization

As another illustration of the consequences of Theorem 1, we now turn to the columnwise $(1,2)$-norm previously defined in Example 2. As before, specializing Theorem 1 to this decomposable regularizer yields the following guarantee:

Corollary 2. *Suppose $\Theta^\star$ has rank at most r with $\|\Theta^\star\|_{\infty,2} \le \frac{\alpha}{\sqrt{d_2}}$, and $\Gamma^\star$ has at most s non-zero columns. If the noise matrix W has i.i.d. $N(0, \nu^2/(d_1 d_2))$ entries, and we solve the convex program (10) with $\lambda_d = \frac{8\nu}{\sqrt{d_1}} + \frac{8\nu}{\sqrt{d_2}}$ and*

$$\mu_d = 8\nu \left(\sqrt{\frac{1}{d_2}} + \sqrt{\frac{\log d_2}{d_1 d_2}}\right) + \frac{4\alpha}{\sqrt{d_2}},$$

then with probability greater than $1 - \exp\left(-2\log(d_2)\right)$, the error $e^2(\widehat{\Theta}, \widehat{\Gamma})$ of any solution is bounded by

$$c_1 \nu^2 \left(\frac{r}{d_1} + \frac{r}{d_2} + \frac{s}{d_2} + \frac{s \log d_2}{d_1 d_2}\right) + c_2 \frac{\alpha^2 s}{d_2}. \qquad (18)$$

Remarks: Note that the setting of λ_d is the same as in Corollary 1, whereas the parameter μ_d is chosen based on upper bounding $\|W\|_{\infty,2}$, corresponding to the dual norm of the columnwise $(1,2)$-norm. As with Corollary 1, an alternative argument can be used to replace the logarithmic term with $\log(d_2/s)$, and the resulting bound can be shown to be minimax optimal.

3.5. Lower Bounds

For the case of i.i.d Gaussian noise matrices, Corollaries 1 and 2 guarantee that our estimators achieve certain Frobenius errors. In this section, we turn to

the complementary question: what are the fundamental (algorithm-independent) limits of accuracy in noisy matrix decomposition? Given some family $\mathcal{F}$ of matrices, the associated minimax error is given by

$$\mathfrak{M}(\mathcal{F}) := \inf_{(\widetilde{\Theta},\widetilde{\Gamma})} \sup_{(\Theta^\star,\Gamma^\star)} \mathbb{E}\left[\|\widetilde{\Theta} - \Theta^\star\|_F^2 + \|\widetilde{\Gamma} - \Gamma^\star\|_F^2\right],$$

where the infimum ranges over all estimators $(\widetilde{\Theta},\widetilde{\Gamma})$ that are (measurable) functions of the data Y, and the supremum ranges over all pairs $(\Theta^\star,\Gamma^\star) \in \mathcal{F}$. Here the expectation is taken over the Gaussian noise matrix W, under the linear observation model (1).

For the case of elementwise sparsity, the relevant family is the set $\mathcal{F}_{\mathrm{sp}}(r,s,\alpha)$ of matrix pairs such that $\Theta^\star$ has rank at most r and elementwise spikiness $\|\Theta^\star\|_\infty \leq \frac{\alpha}{\sqrt{d_1 d_2}}$; and such that $\Gamma^\star$ has at most s non-zero entries. Similarly, for columnwise sparsity, we let $\mathcal{F}_{\mathrm{col}}(r,s,\alpha)$ be the set of matrix pairs such that $\Theta^\star$ has rank at most r and columnwise spikiness $\|\Theta^\star\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}}$; and such that $\Gamma^\star$ has at most s non-zero columns.

Theorem 2 (Minimax lower bound). *There is a universal constant $c_0 > 0$ such that for all $\alpha \geq 8\sqrt{\log(d_1 d_2)}$, the minimax risk $\mathfrak{M}(\mathcal{F}_{\mathrm{sp}}(r,s,\alpha))$ is lower bounded by*

$$c_0 \nu^2 \left(\frac{r}{d_1} + \frac{r}{d_2} + \frac{s \log(\frac{d_1 d_2}{s})}{d_1 d_2} \right) + c_0 \frac{\alpha^2 s}{d_1 d_2},$$

and the minimax risk $\mathfrak{M}(\mathcal{F}_{\mathrm{col}}(r,s,\alpha))$ is lower bounded by

$$c_0 \nu^2 \left(\frac{r}{d_1} + \frac{r}{d_2} + \frac{s}{d_2} + \frac{s \log(\frac{d_2 - s}{s/2})}{d_1 d_2} \right) + c_0 \frac{\alpha^2 s}{d_2}.$$

These lower bounds match the sharper versions of Corollaries 1 and 2 that we discuss in the text following the corollaries up to constants. Consequently, up to constant factors, Corollaries 1 and 2 cannot be improved upon by any estimator.

4. Proof sketches for main theorem corollaries

Due to lack of space, we only outline the proofs of the corollaries, referring the reader to the long version (Agarwal et al., 2011) for all technical details.

4.1. Proof of Corollary 1

In order to prove this result, we need to verify that the stated choice of (λ_d, μ_d) satisfies the requirements of Theorem 1. Recall that here $\mathcal{R}(\cdot) = \|\cdot\|_1$ and $\mathcal{R}^*(\cdot) = \|\cdot\|_\infty$. Based on the discussion following the corollary, we only need to verify that $\lambda_d \geq$

$4\|W\|_{\mathrm{op}}$ and $\mu_d \geq 4\|W\|_\infty + 4\frac{\alpha}{\sqrt{d_1 d_2}}$. By known results on the singular values of Gaussian random matrices (Davidson & Szarek, 2001), for the given Gaussian random matrix, we have

$$\mathbb{P}\left[\|W\|_{\mathrm{op}} \geq 4\nu\left\{\frac{1}{\sqrt{d_1}} + \frac{1}{\sqrt{d_2}}\right\}\right] \leq 2\exp\left(-c(d_1 + d_2)\right).$$

Consequently, setting $\lambda_d \geq 16\nu\{\frac{1}{\sqrt{d_1}} + \frac{1}{\sqrt{d_2}}\}$ ensures that the requirement (14) is satisfied. As for the associated requirement for μ_d, it suffices to upper bound $\|W\|_\infty$. Since the entries of W are i.i.d. and sub-Gaussian with parameter $\nu/\sqrt{d_1 d_2}$, the sub-Gaussian tail bound combined with union bound yields

$$\mathbb{P}\left[\|W\|_\infty \geq 4\frac{\nu}{\sqrt{d_1 d_2}} \log(d_1 d_2)\right] \leq \exp(-\log d_1 d_2),$$

from which the statement of Corollary 1 follows.

In order to obtain the sharper minimax optimal result, need to directly analyze the $|\langle\!\langle W, \widehat{\Delta}^\Gamma \rangle\!\rangle|$ term using a result of Gordon et al. (2007). Details are omitted due to lack of space.

4.2. Proof of Corollary 2

Recall that here $\mathcal{R}(\cdot) = \|\cdot\|_{1,2}$ and $\mathcal{R}^*(\cdot) = \|\cdot\|_{\infty,2}$. To prove the Corollary, we need to show that the conditions of Theorem 1 on λ_d, μ_d hold with high probability. It is easily seen that the setting of λ_d is the same as Corollary 1. Hence, we only need to establish an upper bound on $\|W\|_{\infty,2}$ to complete the proof. Let W_k be the k_{th} column of the matrix. Then for any fixed k

$$\mathbb{P}\left[\|W_k\|_2 \geq \mathbb{E}\|W_k\|_2 + t\right] \leq \exp\left(-\frac{t^2 d_1 d_2}{\nu^2}\right).$$

Since W_k is a Gaussian random vector, we have

$$\mathbb{E}\|W_k\|_2 \leq \frac{\nu}{\sqrt{d_1 d_2}}\sqrt{d_1} = \frac{\nu}{\sqrt{d_2}},$$

and combined with union bound,

$$\mathbb{P}\left[\max_k \|W_k\|_2 \geq \frac{\nu}{\sqrt{d_2}} + t\right] \leq d_2 \exp\left(-\frac{t^2 d_1 d_2}{\nu^2}\right).$$

Setting $t = 2\nu\sqrt{\frac{\log d_2}{d_1 d_2}}$ gives that with probability at least $1 - \exp\left(-3\log d_2\right)$

$$\|W\|_{\infty,2} \leq \frac{\nu}{\sqrt{d_2}} + 2\nu\sqrt{\frac{\log d_2}{d_1 d_2}}.$$

5. Experimental results

In this section, we present simulation studies demonstrating the accuracy of our theory. The first set of experiments apply to a low-rank matrix $\Theta^\star \in \mathbb{R}^{d \times d}$ corrupted by an arbitrary sparse matrix $\Gamma^\star$. For positive parameters γ and β, we varied the rank of $\Theta^\star$ as $r = \gamma d$, and the sparsity of $\Gamma^\star$ as $s = \beta d^2 / \log(d^2)$, and we generated the noise matrix W with i.i.d. $N(0, 1/d)$ entries. With these choices, Corollary 1 guarantees that (w.h.p.) the squared Frobenius error should be upper bounded as $c_1 \gamma + c_2 \beta$, where c_1, c_2 are universal constants. For dimension $d = 100$ and sparsity $s = 2171$, Figure 1 shows plots of squared Frobenius error as the rank parameter γ is varied: as predicted, the error grows linearly with γ.

Figure 2. Plot of the squared Frobenius norm joint error of $\widehat{\Theta}$ and $\widehat{\Gamma}$ where we fix $d = 100$ and $r = 10$, and vary $\beta \in \{0.5 : 0.5 : 5, 6, 7, 10, 20\}$.

Figure 1. Plot of the squared Frobenius norm joint error of $\widehat{\Theta}$ and $\widehat{\Gamma}$. We vary $\gamma \in \{0.05 : 0.05 : 0.5\}$, and set $d = 100$ with sparsity level $s = 2171$. The growth of the function is linear in γ or r, which experimentally demonstrates the $\sqrt{r}$ error scaling predicted by the theory.

For dimension $d = 100$ and rank $r = 10$, Figure 2 shows the squared Frobenius error as the sparsity parameter β is varied. Consistent with the predictions of Corollary 1, the squared Frobenius error scales linearly with β.

We now turn to simulations for a low-rank matrix corrupted by a column-sparse matrix $\Gamma^\star$, in this case studying how the error scales with the matrix dimension d. In all cases, for a matrix $\Theta^\star$ with rank r, we generate the matrix $\Gamma^\star$ with $s = 3r$ non-zero columns of arbitrary magnitude.

Figure 3 plots the squared Frobenius error versus the matrix dimension d; it contains two curves, one with rank $r = 10$ and $d \in \{100 : 25 : 300\}$, and the sec-

ond with rank $r = 15$ and $d \in 1.5 * \{100 : 25 : 300\}$. In both cases, the error decreases as d increases, consistent with Corollary 2. Furthermore, when r is increased by a factor of $3/2$, the dimension needs to be increased by the same factor in order to achieve the same error. In Figure 4, we plot the squared inverse
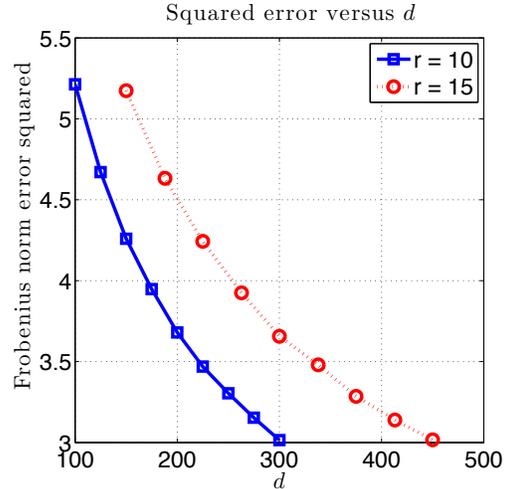
Figure 3. Plot demonstrating that as d increases we see a decrease in the total error as predicted. Here, $s = 3r$ and $r = 10$ and 15.

Frobenius error with dimension and observe that the squared inverse error increases linearly in the dimension d, consistent with Corollary 2.

6. Discussion

In this paper, we provided conditions under which a low-rank matrix and a sparse matrix can be recovered
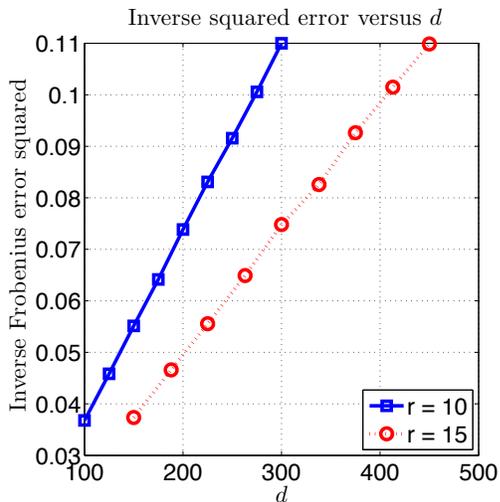
Figure 4. Plot of the inverse squared Frobenius norm error, which demonstrates that the error decreases as $1/\sqrt{d}$ for $s = 3r$ and $r = 10$ and 15. Furthermore, scaling the rank and matrix dimensions by the same amount results in nearly identical errors.

from the noisy sum. The results apply to broad classes of structural assumptions on the low-dimensional component generalizing several previous results to a noisy observation model. The error rates of our estimator can be shown to be minimax optimal for both elementwise and columnwise sparsity models. A key feature of our results is the weakening of the incoherence assumption made in most of the existing literature. In future work, it will be interesting to study other family of decompositions in which recovery is possible under some regularity conditions.

References

Agarwal, A., Negahban, S. N., and Wainwright, M. J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. 2011. URL http://arxiv.org/abs/1012.4807.

Candes, E. J., Li, X., Ma, Y., and Wright, J. Robust Principal Component Analysis? 2009. URL http://arxiv.org/abs/0912.3599.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. Technical report, MIT, June 2009. Available at arXiv:0906.2220v1.

Chandrasekaran, V., Parillo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. 2010. URL http://arxiv.org/abs/1008.1290.

Davidson, K. R. and Szarek, S. J. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pp. 317–336. Elsevier, Amsterdam, NL, 2001.

Gordon, Y., Litvak, A. E., Mendelson, S., and Pajor, A. Gaussian averages of interpolated bodies and applications to approximate reconstruction. *Journal of Approximation Theory*, 149:59–73, 2007.

Hsu, D., Kakade, S. M., and Zhang, T. Robust Matrix Decomposition with Outliers. 2010. URL http://arxiv.org/abs/1011.1518.

McCoy, M. and Tropp, J. Two Proposals for Robust PCA using Semidefinite Programming. 2010. URL http://arxiv.org/abs/1012.1086.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. Technical report, UC Berkeley, August 2010.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS Conference*, Vancouver, Canada, December 2009. Full length version arxiv:1010.2731v1.

Recht, B., Fazel, M., and Parrilo, P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52 (3):471–501, 2010.

Xu, H., Caramanis, C., and Sanghavi, S. Robust PCA via Outlier Pursuit. 2010. URL http://arxiv.org/abs/1010.4237.