The Annals of Statistics
2012, Vol. 40, No. 2, 1171–1197
DOI: 10.1214/12-AOS1000

NOISY MATRIX DECOMPOSITION VIA CONVEX RELAXATION: OPTIMAL RATES IN HIGH DIMENSIONS[1]

BY ALEKH AGARWAL[2], SAHAND NEGAHBAN[3] AND MARTIN J. WAINWRIGHT[3]

University of California, Berkeley, Massachusetts Institute of Technology and University of California, Berkeley

We analyze a class of estimators based on convex relaxation for solving high-dimensional matrix decomposition problems. The observations are noisy realizations of a linear transformation $\mathfrak{X}$ 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; this set-up includes many statistical models of interest, including factor analysis, multi-task regression and robust covariance estimation. We derive a general theorem that bounds the Frobenius norm error for an estimate of the pair $(\Theta^\star, \Gamma^\star)$ obtained by solving a convex optimization problem that combines the nuclear norm with a general decomposable regularizer. Our results use a "spikiness" condition that is related to, but milder than, singular vector incoherence. We specialize our general result to two cases that have been studied in past work: low rank plus an entrywise sparse matrix, and low rank plus a columnwise sparse matrix. For both models, our theory yields nonasymptotic Frobenius error bounds for both deterministic and stochastic noise matrices, and applies to matrices $\Theta^\star$ that can be exactly or approximately low rank, and matrices $\Gamma^\star$ that can be exactly or approximately sparse. Moreover, for the case of stochastic noise matrices and the identity observation operator, we establish matching lower bounds on the minimax error. The sharpness of our nonasymptotic predictions is confirmed by numerical simulations.

1. Introduction. The focus of this paper is a class of high-dimensional matrix decomposition problems of the following variety. 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, either exactly or approximately, and allows for a class of low-dimensional structures for

Received March 2012.
[1]Supported in part by AFOSR Grant 09NL184.
[2]Supported in part by a Microsoft Research Fellowship and a Google Ph.D. Fellowship.
[3]Supported in part by NSF Grant CDI-0941742.
MSC2010 subject classifications. Primary 62F30, 62F30; secondary 62H12.
Key words and phrases. High-dimensional inference, nuclear norm, composite regularizers.

the second component $\Gamma^\star$. Two particular cases of structure for $\Gamma^\star$ that have been considered in past work are elementwise sparsity [7–9] and column-wise sparsity [13, 21].

Problems of matrix decomposition are motivated by a variety of applications. Many classical methods for dimensionality reduction, among them factor analysis and principal components analysis (PCA), are based on estimating a low-rank matrix from data. Different forms of robust PCA can be formulated in terms of matrix decomposition using the matrix $\Gamma^\star$ to model the gross errors [7, 9, 21]. Similarly, certain problems of robust covariance estimation can be described using matrix decompositions with a column/row-sparse structure, as we describe in this paper. The problem of low rank plus sparse matrix decomposition also arises in Gaussian covariance selection with hidden variables [8]. Matrix decompositions also arise in multi-task regression [15, 20, 22], which involve solving a collection of regression problems, referred to as tasks, over a common set of features. For some features, one expects their weighting to be preserved across tasks, which can be modeled by a low-rank constraint, whereas other features are expected to vary across tasks, which can be modeled by a sparse component [3, 5]. See Section 2.1 for further discussion of these motivating applications.

In this paper, we study a noisy linear observation model that captures a number of applications in a unified way. Let $\mathfrak{X}$ be a linear operator that maps matrices in $\mathbb{R}^{d_1 \times d_2}$ to matrices in $\mathbb{R}^{n_1 \times n_2}$. In the simplest case, this observation operator is simply the identity mapping, so that we necessarily have $n_1 = d_1$ and $n_2 = d_2$. However, as we discuss in the sequel, it is useful for certain applications, such as multi-task regression, to consider more general linear operators of this form. Hence, we study the problem of matrix decomposition for the general linear observation model

$$Y = \mathfrak{X}(\Theta^\star + \Gamma^\star) + W, \tag{1}$$

where $\Theta^\star$ and $\Gamma^\star$ are unknown $d_1 \times d_2$ matrices, and $W \in \mathbb{R}^{n_1 \times n_2}$ is some type of observation noise; it 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. As we discuss in Section 2.1, a variety of interesting statistical models can be formulated as instances of the observation model (1). Such models include versions of factor analysis involving nonidentity noise matrices, robust covariance estimation, and multi-task regression with some features shared across tasks, and a sparse subset differing across tasks. Given this set-up, our goal is to recover accurate estimates 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 $\mathcal{R}$.

Most past work on model (1) has focused on the noiseless setting ($W = 0$), and for the identity observation operator [i.e., for which we have $\mathfrak{X}(\Theta^\star + \Gamma^\star) =$

$\Theta^\star + \Gamma^\star]$. Chandrasekaran et al. [9] studied the case when $\Gamma^\star$ is assumed to sparse, with a relatively small number $s \ll d_1 d_2$ of nonzero entries. In the noiseless setting, they gave sufficient conditions for exact recovery for an adversarial sparsity model, meaning the nonzero positions of $\Gamma^\star$ can be arbitrary. Subsequent work by Candes et al. [7] analyzed the same model but under an assumption of random sparsity, meaning that the nonzero positions are chosen uniformly at random. In recent work, Xu et al. [21] have analyzed a different model, in which the matrix $\Gamma^\star$ is assumed to be columnwise sparse, with a relatively small number $s \ll d_2$ of nonzero columns. Their analysis guaranteed approximate recovery for the low-rank matrix. After initial posting of this work, we became aware of recent work by Hsu et al. [11], who derived Frobenius norm error bounds for the case of exact elementwise sparsity. As we discuss in Section 3.4, in this special case, our bounds are based on milder conditions, and yield sharper rates for problems where the rank and sparsity scale with the dimension.

Our main contribution is to provide a general oracle-type result (Theorem 1) on approximate recovery of the unknown decomposition from noisy observations, valid for structural constraints on $\Gamma^\star$ imposed via a decomposable regularizer. The class of decomposable regularizers, introduced in past work by Negahban et al. [14], includes the elementwise ℓ_1-norm and columnwise $(2, 1)$-norm as special cases, as well as various other regularizers used in practice. Our main result is stated in Theorem 1: it provides finite-sample guarantees for estimates obtained by solving a class of convex programs formed using a composite regularizer. We then specialize Theorem 1 to the case of elementwise or columnwise 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. We provide nonasymptotic error bounds for general noise matrices W both for elementwise and columnwise sparse models; see Corollaries 1–6. To the best of our knowledge, these are the first results that apply to this broad class of models, allowing for stochastic as well as deterministic noise, matrix components that are only approximately low-rank and/or sparse, and a general observation operator $\mathfrak{X}$.

In addition, the error bounds obtained by our analysis are sharp, and cannot be improved in general. More precisely, for the case of stochastic noise matrices and the identity observation operator, we prove that the squared Frobenius errors of our estimators are minimax-optimal; see Theorem 2. An interesting feature of our analysis is that, in contrast to previous work [7, 9, 21], we do *not* impose incoherence conditions on the singular vectors of $\Theta^\star$; rather, we control the interaction with a milder condition 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, and has proven useful in the related setting of noisy matrix completion [16]. This constraint does not guarantee identifiability of the models (and hence exact recovery in the noiseless setting), but it does provide a bound on the degree of nonidentifiability. We show that this same term arises in

both the upper and lower bounds on the problem of 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 on achievability, as well as its various corollaries for special cases of the matrix decomposition problem. We complement these achievable results with matching minimax lower bounds in Section 4, and we conclude with a discussion in Section 5. The supplementary material [1] contains numerical simulations that illustrate the sharpness of our theoretical predictions (Section 6), as well as the technical proofs in Section 7.

Notation. For the reader's convenience, we summarize here some of the standard notation used throughout this paper. For any matrix $M \in \mathbb{R}^{d_1 \times d_2}$, we define the *Frobenius norm* $|||M|||_{\mathrm{F}} := \sqrt{\sum_{j=1}^{d_1} \sum_{k=1}^{d_2} M_{jk}^2}$, corresponding to the ordinary Euclidean norm of its entries. We denote its singular values by $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_d(M) \geq 0$, where $d = \min\{d_1, d_2\}$. Its *nuclear norm* is given by $|||M|||_{\mathrm{N}} = \sum_{j=1}^{d} \sigma_j(M)$.

2. Convex relaxations and matrix decomposition. In this paper, we consider a family of regularizers formed by a combination of the nuclear norm $|||\cdot|||_{\mathrm{N}}$, which acts as a convex surrogate to a rank constraint for $\Theta^\star$ (e.g., see Recht et al. [18] 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 decomposable regularizers [14], 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 $(2, 1)$-norm; see Examples 4 and 5, below.

2.1. *Some motivating applications*. We begin with some motivating applications for the general linear observation model with noise (1).

EXAMPLE 1 (Factor analysis with sparse noise). In factor analysis, we have i.i.d. random vectors $Z_i \in \mathbb{R}^d$ assumed to be generated from the model

$$Z_i = LU_i + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, n, \tag{2}$$

where $L \in \mathbb{R}^{d_1 \times r}$ is a loading matrix, and the vectors $U_i \sim N(0, I_{r \times r})$ and $\varepsilon_i \sim N(0, \Gamma^\star)$ are independent. Given n i.i.d. samples from model (2), the goal is to estimate the loading matrix L, or the matrix LL^T, that projects onto column span of L. A simple calculation shows that the covariance matrix of Z_i has the form $\Sigma = LL^T + \Gamma^\star$. Consequently, in the special case when $\Gamma^\star = \sigma^2 I_{d \times d}$, then the range of L is spanned by the top r eigenvectors of Σ, and so we can recover it via standard principal components analysis.

In other applications, we might no longer be guaranteed that $\Gamma^\star$ is the identity, in which case the top r eigenvectors of Σ need not be close to column

span of L. Nonetheless, when $\Gamma^\star$ is a sparse matrix, the problem of estimating LL^T can be understood as an instance of our general observation model (1) with $d_1 = d_2 = d$, and the identity observation operator $\mathfrak{X}$ (so that $n_1 = n_2 = d$). In particular, if the let the observation matrix $Y \in \mathbb{R}^{d\times d}$ be the sample covariance matrix $\frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$, then some algebra shows that $Y = \Theta^\star + \Gamma^\star + W$, where $\Theta^\star = LL^T$ is of rank r, and the random matrix W is re-centered Wishart noise [2]—in particular, the zero-mean matrix

$$W := \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T - \{LL^T + \Gamma^\star\}. \tag{3}$$

When $\Gamma^\star$ is assumed to be sparse, then this constraint can be enforced via the elementwise ℓ_1-norm; see Example 4 to follow.

EXAMPLE 2 (Multi-task regression). Suppose that we are given a collection of d_2 regression problems in $\mathbb{R}^{d_1}$, each of the form $y_j = X\beta_j^* + w_j$ for $j = 1, 2, \ldots, d_2$. Here each $\beta_j^* \in \mathbb{R}^{d_1}$ is an unknown regression vector, $w_j \in \mathbb{R}^n$ is observation noise, and $X \in \mathbb{R}^{n\times d_1}$ is the design matrix. This family of models can be written in a convenient matrix form as $Y = XB^* + W$, where $Y = [y_1 \ \cdots \ y_{d_2}]$ and $W = [w_1 \ \cdots \ w_{d_2}]$ are both matrices in $\mathbb{R}^{n\times d_2}$ and $B^* := [\beta_1^* \ \cdots \ \beta_{d_2}^*] \in \mathbb{R}^{d_1\times d_2}$ is a matrix of regression vectors. Following standard terminology in multi-task learning, we refer to each column of B^* as a *task*, and each row of B^* as a *feature*.

In many applications, it is natural to assume that the feature weightings—that is, the vectors $\beta_j^* \in \mathbb{R}^{d_2}$—exhibit some degree of shared structure across tasks [3, 15, 20, 22]. This type of shared structure can be modeled by imposing a low-rank structure; for instance, in the extreme case of rank one, it would enforce that each β_j^* is a multiple of some common underlying vector. However, many multi-task learning problems exhibit more complicated structures, in which some subset of features are shared across tasks, and some other subset of features vary substantially across tasks [3, 4]. For instance, in the Amazon recommendation system, tasks correspond to different classes of products, such as books, electronics and so on, and features include ratings by users. Some ratings (such as numerical scores) should have a meaning that is preserved across tasks, whereas other features (e.g., the label "boring") are very meaningful in applications to some categories (e.g., books) but less so in others (e.g., electronics).

This kind of structure can be captured by assuming that the unknown regression matrix B^* has a low-rank plus sparse decomposition—namely, $B^* = \Theta^\star + \Gamma^\star$ where $\Theta^\star$ is low-rank and $\Gamma^\star$ is sparse, with a relatively small number of nonzero entries, corresponding to feature/task pairs that differ significantly from the baseline. A variant of this model is based on instead assuming that $\Gamma^\star$ is row-sparse, with a small number of nonzero rows. (In Example 5, to follow, we discuss an appropriate regularizer for enforcing such row or column sparsity.) With this

model structure, we then define the observation operator $\mathfrak{X}: \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{n \times d_2}$ via $A \mapsto XA$, so that $n_1 = n$ and $n_2 = d_2$ in our general notation. In this way, we obtain another instance of the linear observation model (1).

EXAMPLE 3 (Robust covariance estimation). For $i = 1, 2, \ldots, n$, let $U_i \in \mathbb{R}^d$ be samples from a zero-mean distribution with unknown covariance matrix $\Theta^\star$. When the vectors U_i are observed without any form of corruption, then it is straightforward to estimate $\Theta^\star$ by performing PCA on the sample covariance matrix. Imagining that $j \in \{1, 2, \ldots, d\}$ indexes different individuals in the population, now suppose that the data associated with some subset S of individuals is arbitrarily corrupted. This adversarial corruption can be modeled by assuming that we observe the vectors $Z_i = U_i + v_i$ for $i = 1, \ldots, n$, where each $v_i \in \mathbb{R}^d$ is a vector supported on the subset S. Letting $Y = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$ be the sample covariance matrix of the corrupted samples, some algebra shows that it can be decomposed as $Y = \Theta^\star + \Delta + W$, where $W := \frac{1}{n}\sum_{i=1}^n U_i U_i^T - \Theta^\star$ is again a type of re-centered Wishart noise, and the remaining term can be written as

$$\Delta := \frac{1}{n}\sum_{i=1}^n v_i v_i^T + \frac{1}{n}\sum_{i=1}^n (U_i v_i^T + v_i U_i^T). \tag{4}$$

Note that Δ itself is not a column-sparse or row-sparse matrix; however, since each vector $v_i \in \mathbb{R}^d$ is supported only on some subset $S \subset \{1, 2, \ldots, d\}$, we can write $\Delta = \Gamma^\star + (\Gamma^\star)^T$, where $\Gamma^\star$ is a column-sparse matrix with entries only in columns indexed by S. This structure can be enforced by the use of the column-sparse regularizer (12), as described in Example 5 to follow.

2.2. *Convex relaxation for noisy matrix decomposition.* Given the observation model $Y = \mathfrak{X}(\Theta^\star + \Gamma^\star) + W$, it is natural to consider an estimator based on solving the regularized least-squares program

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

Here (λ_d, μ_d) are nonnegative regularizer parameters, to be chosen by the user. Our theory also provides choices of these parameters that guarantee good properties of the associated estimator. Although this 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 constraints, model (1) is unidentifiable, even in the noiseless setting ($W = 0$). Indeed, as discussed in past work [7, 9, 21], no method can recover the components $(\Theta^\star, \Gamma^\star)$ unless the low-rank component is "incoherent" with the matrix $\Gamma^\star$. For instance, taking $\Gamma^\star$ to be 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 [7, 9, 21] has ruled out these 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; see the discussion following Example 4 for more details.

In this paper, we impose a related but milder condition, previously introduced in our past work on matrix completion [16], with the goal of performing approximate recovery. To be clear, this condition does not guarantee identifiability, but rather provides a bound on the *radius of nonidentifiability*. It should be noted that nonidentifiability is a feature common to many high-dimensional statistical models.[4] Moreover, 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. Indeed, one of our main contributions is to establish minimax-optimality of our rates, meaning that no algorithm can be substantially better over the matrix classes that we consider.

For a given regularizer $\mathcal{R}$, we define the quantity $\kappa_d(\mathcal{R}) := \sup_{V \neq 0} |||V|||_{\mathrm{F}} / \mathcal{R}(V)$, which measures the relation between the regularizer and the Frobenius norm. Moreover, we define the associated dual norm

$$\mathcal{R}^*(U) := \sup_{\mathcal{R}(V) \leq 1} \langle\langle V, U \rangle\rangle, \tag{5}$$

where $\langle\langle V, U \rangle\rangle := \operatorname{trace}(V^T U)$ 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 and $\Gamma^\star$ via the quantity

$$\varphi_{\mathcal{R}}(\Theta) := \kappa_d(\mathcal{R}^*) \mathcal{R}^*(\Theta). \tag{6}$$

More specifically, we analyze the family of estimators

$$\min_{(\Theta, \Gamma)} \left\{ \frac{1}{2} |||Y - \mathfrak{X}(\Theta + \Gamma)|||_{\mathrm{F}}^2 + \lambda_d |||\Theta|||_{\mathrm{N}} + \mu_d \mathcal{R}(\Gamma) \right\}, \tag{7}$$

subject to $\varphi_{\mathcal{R}}(\Theta) \leq \alpha$ for some fixed parameter α.

2.3. *Some examples*. Let us consider some examples to discuss specific forms of the estimator (7), and the role of the additional constraint.

EXAMPLE 4 (Sparsity and elementwise ℓ_1-norm). Suppose that $\Gamma^\star$ is assumed to be sparse, with $s \ll d_1 d_2$ nonzero entries. In this case, the sum $\Theta^\star + \Gamma^\star$ corresponds to the sum of a low rank matrix with a sparse matrix. Motivating applications include the problem of factor analysis (Example 1), as well as certain formulations of robust PCA [7] and model selection in Gauss–Markov random

[4] For instance, see the paper [17] for discussion of nonidentifiability in high-dimensional sparse regression.

fields with hidden variables [8]. Given the sparsity of $\Gamma^\star$, an appropriate choice of regularizer is the elementwise ℓ_1-norm

$$\mathcal{R}(\Gamma) = \|\Gamma\|_1 := \sum_{j=1}^{d_1}\sum_{k=1}^{d_2}|\Gamma_{jk}|. \tag{8}$$

With this choice, it is straightforward to verify that

$$\mathcal{R}^*(Z) = \|Z\|_\infty := \max_{j=1,\ldots,d_1}\max_{k=1,\ldots,d_2}|Z_{jk}|, \tag{9}$$

and moreover, that $\kappa_d(\mathcal{R}^*) = \sqrt{d_1 d_2}$. Consequently, in this specific case, the general convex program (7) takes the form

$$\min_{(\Theta,\Gamma)}\left\{\frac{1}{2}|||Y - \mathfrak{X}(\Theta+\Gamma)|||_{\mathrm{F}}^2 + \lambda_d|||\Theta|||_{\mathrm{N}} + \mu_d\|\Gamma\|_1\right\}, \tag{10}$$

subject to $\|\Theta\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}$. 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. Indeed, this type of spikiness control has proven useful in analysis of nuclear norm relaxations for noisy matrix completion [16]. To gain intuition for the parameter α, if we consider matrices with $|||\Theta|||_{\mathrm{F}} \approx 1$, as is appropriate to keep a constant signal-to-noise ratio in the noisy model (1), then setting $\alpha \approx 1$ allows only for matrices for which $|\Theta_{jk}| \approx 1/\sqrt{d_1 d_2}$ in all entries. In contrast, the maximally spiky matrix with all its mass in a single position, requires $\alpha \approx \sqrt{d_1 d_2}$. In practice, we are interested in settings of α lying between these two extremes.

Past work on ℓ_1-forms of matrix decomposition has imposed singular vector incoherence conditions that are related to, but different from, our spikiness condition. More concretely, if we write the SVD of the low-rank component as $\Theta^\star = UDV^T$ where D is diagonal, and $U \in \mathbb{R}^{d_1\times r}$ and $V \in \mathbb{R}^{d_2\times r}$ are matrices of the left and right singular vectors. Singular vector incoherence bounds quantities such as

$$\left\|UU^T - \frac{r}{d_1}I_{d_1\times d_1}\right\|_\infty, \qquad \left\|VV^T - \frac{r}{d_2}I_{d_2\times d_2}\right\|_\infty \quad \text{and} \quad \|UV^T\|_\infty \tag{11}$$

all of which measure the degree of "coherence" between the singular vectors and the canonical basis. A remarkable feature of such conditions is that they have no dependence on the *singular values* of $\Theta^\star$. This lack of dependence makes sense in the noiseless setting, where exact recovery is the goal. For noisy models, in contrast, one should only be concerned with recovering components with "large" singular values. In this context, our bound on the maximum element $\|\Theta^\star\|_\infty$, or equivalently on the quantity $\|UDV^T\|_\infty$, is natural. Note that it imposes no constraint on the matrices UU^T or VV^T, and moreover it uses the diagonal matrix of singular values as a weight in the ℓ_∞ bound. Moreover, we note that there are many matrices for which $\|\Theta^\star\|_\infty$ satisfies a reasonable bound, whereas the incoherence measures are poorly behaved; for example, see Section 3.4.2 in the paper [16].

EXAMPLE 5 (Column-sparsity and block columnwise regularization). Other applications involve models in which $\Gamma^\star$ has a relatively small number $s \ll d_2$ of nonzero columns (or a relatively small number $s \ll d_1$ of nonzero rows). Such applications include the multi-task regression problem from Example 2, the robust covariance problem from Example 3 as well as a form of robust PCA considered by Xu et al. [21]. In this case, it is natural to constrain Γ via the $(2,1)$-norm regularizer

$$\mathcal{R}(\Gamma) = \|\Gamma\|_{2,1} := \sum_{k=1}^{d_2} \|\Gamma_k\|_2, \tag{12}$$

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

$$\mathcal{R}^*(U) = \|U\|_{2,\infty} := \max_{k=1,2,\ldots,d_2} \|U_k\|_2, \tag{13}$$

where U_k denotes the kth column of U, and that $\kappa_d(\mathcal{R}^*) = \sqrt{d_2}$. Consequently, in this specific case, the general convex program (7) takes the form

$$\min_{(\Theta,\Gamma)} \left\{ \frac{1}{2} |||Y - \mathfrak{X}(\Theta+\Gamma)|||_{\mathrm{F}}^2 + \lambda_d |||\Theta|||_{\mathrm{N}} + \mu_d \|\Gamma\|_{2,1} \right\}, \tag{14}$$

subject to $\|\Theta\|_{2,\infty} \leq \frac{\alpha}{\sqrt{d_2}}$. As before, the constraint $\|\Theta\|_{2,\infty}$ serves to limit the "spikiness" of the low rank component, where in this case, spikiness is measured in a columnwise manner. Again, it is natural to consider matrices such that $|||\Theta^\star|||_{\mathrm{F}} \approx 1$, so that the signal-to-noise ratio in the observation model (1) stays fixed. Thus, if $\alpha \approx 1$, then we are restricted to matrices for which $\|\Theta^\star_k\|_2 \approx \frac{1}{\sqrt{d_2}}$ for all columns $k = 1, 2, \ldots, d_2$. At the other extreme, in order to permit a maximally "column-spiky" matrix (i.e., with a single nonzero column of ℓ_2-norm roughly 1), we need to set $\alpha \approx \sqrt{d_2}$. As before, of practical interest are settings of α lying between these two extremes.

3. Main results and their consequences. In this section, we state our main results and discuss some of their consequences. Our first result applies to the family of convex programs (7) whenever $\mathcal{R}$ belongs to the class of decomposable regularizers, and the least-squares loss associated with the observation model satisfies a specific form of restricted strong convexity [14]. We begin in Section 3.1 by defining the notion of decomposability, and then illustrating how both the elementwise-ℓ_1 and columnwise-$(2,1)$-norms are instances of decomposable regularizers. In Section 3.2, we define the form of restricted strong convexity appropriate to our setting. Section 3.3 contains the statement of our main result about the M-estimator (7), while Sections 3.4 and 3.6 are devoted to its consequences for the cases of elementwise sparsity and columnwise sparsity, respectively. In Section 3.5, we complement our analysis of the convex program (7) by showing that, in the special case of the identity operator, a simple two-step method can achieve similar rates

(up to constant factors). We also provide an example showing that the two-step method can fail for more general observation operators. Matching lower bounds on the minimax errors in the case of the identity operator and Gaussian noise are presented in Section 4 to follow.

3.1. *Decomposable regularizers.* 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 that is sufficient to cover the examples of interest in this paper:

DEFINITION 1. Given a subspace $\mathbb{M} \subseteq \mathbb{R}^{d_1 \times d_2}$ and its orthogonal complement $\mathbb{M}^\perp$, a norm $\mathcal{R}$ is *decomposable with respect to* $(\mathbb{M}, \mathbb{M}^\perp)$ if

$$\mathcal{R}(U + V) = \mathcal{R}(U) + \mathcal{R}(V) \qquad \text{for all } U \in \mathbb{M} \text{ and } V \in \mathbb{M}^\perp. \tag{15}$$

To provide some intuition, the subspace $\mathbb{M}$ 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 $\mathbb{M}$. The orthogonal complement $\mathbb{M}^\perp$ represents deviations away from the model subspace, and equality (15) guarantees that such deviations are penalized as much as possible.

As discussed at more length in Negahban et al. [14], a large class of norms is decomposable with respect to interesting[5] subspace pairs. Of particular relevance to us is the decomposability of the elementwise ℓ_1-norm $\|\Gamma\|_1$ and the columnwise $(2, 1)$-norm $\|\Gamma\|_{2,1}$, as discussed in Examples 4 and 5, respectively.

Decomposability of $\mathcal{R}(\cdot) = \|\cdot\|_1$. 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

$$\mathbb{M}(S) := \{U \in \mathbb{R}^{d_1 \times d_2} \mid U_{jk} = 0 \text{ for all } (j, k) \notin S\} \tag{16}$$

and $\mathbb{M}^\perp(S) := (\mathbb{M}(S))^\perp$. It is easy to see that for any pair of matrices $U \in \mathbb{M}(S)$ and $U' \in \mathbb{M}^\perp(S)$, we have the splitting $\|U + U'\|_1 = \|U\|_1 + \|U'\|_1$, showing that the ℓ_1-norm is decomposable with respect to the subspace pair $(\mathbb{M}(S), \mathbb{M}^\perp(S))$.

Decomposability of $\mathcal{R}(\cdot) = \|\cdot\|_{2,1}$. Similarly, the columnwise $(2, 1)$-norm is also decomposable with respect to appropriately defined subspaces, indexed by subsets $C \subseteq \{1, 2, \ldots, d_2\}$ of column indices. Indeed, using V_k to denote the kth column of the matrix V, define

$$\mathbb{M}(C) := \{V \in \mathbb{R}^{d_1 \times d_2} \mid V_k = 0 \text{ for all } k \notin C\} \tag{17}$$

[5]Note that any norm is decomposable wrt the pair $(\mathbb{M}, \mathbb{M}^\perp) = (\mathbb{R}^{d_1 \times d_2}, \{0\})$.

and $\mathbb{M}^{\perp}(C) := (\mathbb{M}(C))^{\perp}$. Again, it is easy to check that for any pair $V \in \mathbb{M}(C)$, $V' \in \mathbb{M}^{\perp}(C)$, we have $\|V + V'\|_{2,1} = \|V\|_{2,1} + \|V'\|_{2,1}$, thus verifying the decomposability property.

For any decomposable regularizer and subspace $\mathbb{M} \neq \{0\}$, we define the compatibility constant

$$\Psi(\mathbb{M}, \mathcal{R}) := \sup_{U \in \mathbb{M}, U \neq 0} \frac{\mathcal{R}(U)}{|||U|||_{\mathrm{F}}}. \tag{18}$$

This quantity measures the compatibility between the Frobenius norm and the regularizer over the subspace $\mathbb{M}$. For example, for the ℓ_1-norm and the set $\mathbb{M}(S)$ previously defined (16), it is straightforward to show that $\Psi(\mathbb{M}(S); \|\cdot\|_1) = \sqrt{s}$.

3.2. *Restricted strong convexity.* Given a loss function, the general notion of strong convexity involves establishing a quadratic lower bound on the error in the first-order Taylor approximation [6]. In our setting, the loss is the quadratic function $\mathcal{L}(\Omega) = \frac{1}{2}|||Y - \mathfrak{X}(\Omega)|||_{\mathrm{F}}^2$ (where we use $\Omega = \Theta + \Gamma$), so that the first-order Taylor series error at Ω in the direction of the matrix Δ is given by

$$\mathcal{L}(\Omega + \Delta) - \mathcal{L}(\Omega) - \nabla\mathcal{L}(\Omega)^T \Delta = \tfrac{1}{2}|||\mathfrak{X}(\Delta)|||_{\mathrm{F}}^2. \tag{19}$$

Consequently, strong convexity is equivalent to a lower bound of the form $\frac{1}{2}\|\mathfrak{X}(\Delta)\|_2^2 \geq \frac{\gamma}{2}|||\Delta|||_{\mathrm{F}}^2$, where $\gamma > 0$ is the strong convexity constant.

Restricted strong convexity is a weaker condition that also involves a norm defined by the regularizers. In our case, for any pair (μ_d, λ_d) of positive numbers, we first define the weighted combination of the two regularizers

$$\mathcal{Q}(\Theta, \Gamma) := |||\Theta|||_{\mathrm{N}} + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma). \tag{20}$$

For a given matrix Δ, we can use this weighted combination to define an associated norm

$$\Phi(\Delta) := \inf_{\Theta + \Gamma = \Delta} \mathcal{Q}(\Theta, \Gamma), \tag{21}$$

corresponding to the minimum of $\mathcal{Q}(\Theta, \Gamma)$ over all decompositions[6] of Δ.

DEFINITION 2 (RSC). The quadratic loss with linear observation operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{n_1 \times n_2}$ satisfies restricted strong convexity with respect to the norm Φ and with parameters (γ, τ_n) if

$$\frac{1}{2}|||\mathfrak{X}(\Delta)|||_{\mathrm{F}}^2 \geq \frac{\gamma}{2}|||\Delta|||_{\mathrm{F}}^2 - \tau_n \Phi^2(\Delta) \qquad \text{for all } \Delta \in \mathbb{R}^{d_1 \times d_2}. \tag{22}$$

[6]Defined this way, $\Phi(\Delta)$ is the infimal-convolution of the two norms $|||\cdot|||_{\mathrm{N}}$ and $\mathcal{R}$, which is a very well studied object in convex analysis (see, e.g., [19]).

Note that if condition (22) holds with $\tau_n = 0$ and any $\gamma > 0$, then we recover the usual definition of strong convexity (with respect to the Frobenius norm). In the special case of the identity operator [i.e., $\mathfrak{X}(\Theta) = \Theta$], such strong convexity does hold with $\gamma = 1$. More general observation operators require different choices of the parameter γ, as well as $\tau_n > 0$.

While RSC establishes a form of (approximate) identifiability in general, here the error Δ is a combination of the error in estimating $\Theta^\star$ and $\Gamma^\star$, which we denote by Δ^Θ and Δ^Γ, respectively. Consequently, we require a further lower bound on $|||\Delta|||_{\mathrm{F}}$ in terms of $|||\Delta^\Theta|||_{\mathrm{F}}$ and $|||\Delta^\Gamma|||_{\mathrm{F}}$ in the proof of our main results so as to guarantee (approximate) identifiability.

3.3. *Results for general regularizers and noise.* We begin by stating a result for a general observation operator $\mathfrak{X}$, a general decomposable regularizer $\mathcal{R}$ and a general noise matrix W. In later subsections, we specialize this result to particular choices of observation operator, regularizers, and stochastic noise matrices. In all our results, we measure error using the *squared Frobenius norm* summed across both matrices

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) := |||\widehat{\Theta} - \Theta^\star|||_{\mathrm{F}}^2 + |||\widehat{\Gamma} - \Gamma^\star|||_{\mathrm{F}}^2. \tag{23}$$

With this notation, the following result applies to the observation model $Y = \mathfrak{X}(\Gamma^\star + \Theta^\star) + W$, where the low-rank matrix satisfies $\varphi_{\mathcal{R}}(\Theta^\star) \le \alpha$. Our upper bound on the squared Frobenius error consists of three terms,

$$\mathcal{K}_\Theta^\star := \frac{\lambda_d^2}{\gamma^2}\left\{ r + \frac{\gamma}{\lambda_d} \sum_{j=r+1}^{d} \sigma_j(\Theta^\star) \right\}, \tag{24a}$$

$$\mathcal{K}_\Gamma^\star := \frac{\mu_d^2}{\gamma^2}\left\{ \Psi^2(\mathbb{M}; \mathcal{R}) + \frac{\gamma}{\mu_d} \mathcal{R}(\Pi_{\mathbb{M}^\perp}(\Gamma^\star)) \right\}, \tag{24b}$$

$$\mathcal{K}_{\tau_n} := \frac{\tau_n}{\gamma}\left\{ \sum_{j=r+1}^{d} \sigma_j(\Theta^\star) + \frac{\mu_d}{\lambda_d} \mathcal{R}(\Pi_{\mathbb{M}^\perp}(\Gamma^\star)) \right\}^2. \tag{24c}$$

As will be clarified shortly, these terms correspond to the errors associated with the low-rank term ($\mathcal{K}_\Theta^\star$), the sparse term ($\mathcal{K}_\Gamma^\star$) and additional error ($\mathcal{K}_{\tau_n}$), due to $\tau_n > 0$ in the RSC condition (22).

THEOREM 1. *Suppose that the observation operator $\mathfrak{X}$ satisfies the RSC condition (22) with curvature $\gamma > 0$, and a tolerance τ_n such that there exists an integer $r \in 1, 2, \ldots, \min\{d_1, d_2\}$, for which*

$$128\tau_n r < \frac{\gamma}{4} \quad \textit{and} \quad 64\tau_n \left(\Psi(\mathbb{M}; \mathcal{R}) \frac{\mu_d}{\lambda_d} \right)^2 < \frac{\gamma}{4}. \tag{25}$$

Then if we solve the convex program (7) *with regularization parameters* (λ_d, μ_d) *satisfying*

$$\lambda_d \geq 4|||\mathfrak{X}^*(W)|||_{\mathrm{op}} \quad \text{and} \quad \mu_d \geq 4\mathcal{R}^*(\mathfrak{X}^*(W)) + \frac{4\gamma\alpha}{\kappa_d}, \tag{26}$$

there are universal constant c_j, $j = 1, 2, 3$ *such that for any matrix pair* $(\Theta^\star, \Gamma^\star)$ *satisfying* $\varphi_{\mathcal{R}}(\Theta^\star) \leq \alpha$ *and any* $\mathcal{R}$*-decomposable pair* $(\mathbb{M}, \mathbb{M}^\perp)$, *any optimal solution* $(\widehat{\Theta}, \widehat{\Gamma})$ *satisfies*

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) \leq c_1\mathcal{K}^\star_\Theta + c_2\mathcal{K}^\star_\Gamma + c_3\mathcal{K}_{\tau_n}. \tag{27}$$

Let us make a few remarks in order to interpret the meaning of this claim.

Deterministic guarantee. To be clear, Theorem 1 is a deterministic statement that applies to any optimum of the convex program (7). Moreover, it actually provides a whole family of upper bounds, one for each choice of the rank parameter r and each choice of the subspace pair $(\mathbb{M}, \mathbb{M}^\perp)$. In practice, these choices are optimized so as to obtain the tightest possible upper bound. As for condition (25), it will be satisfied for a sufficiently large sample size n as long as $\gamma > 0$, and the tolerance τ_n decreases to zero with the sample size. In many cases of interest—including the identity observation operator and multi-task cases—the RSC condition holds with $\tau_n = 0$, so that condition (25) holds as long as $\gamma > 0$.

Interpretation of different terms. Let us focus first on the term $\mathcal{K}^\star_\Theta$, which corresponds to the complexity of estimating the low-rank component. It is further sub-divided into two terms, with the term $\lambda_d^2 r$ corresponding to the *estimation error* associated with a rank r matrix, whereas the term $\lambda_d \sum_{j=r+1}^d \sigma_j(\Theta^\star)$ 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 associated with $\Gamma^\star$, the first of which corresponds to a form of estimation error, whereas the second corresponds to a form of approximation error.

A family of upper bounds. Since inequality (27) corresponds to a family of upper bounds indexed by r and the subspace $\mathbb{M}$, 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 RSC conditions hold with tolerance $\tau_n = 0$, the matrix $\Theta^\star$ is exactly low rank (say rank r), and $\Gamma^\star$ lies within a $\mathcal{R}$-decomposable subspace $\mathbb{M}$. In this case, the approximation errors vanish, and Theorem 1 guarantees that the squared Frobenius error is at most

$$e^2(\widehat{\Theta}; \widehat{\Gamma}) \precsim \lambda_d^2 r + \mu_d^2 \Psi^2(\mathbb{M}; \mathcal{R}), \tag{28}$$

where the $\precsim$ notation indicates that we ignore constant factors.

3.4. *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 is the elementwise ℓ_1-norm, which is decomposable with respect to subspaces of the form (16).

COROLLARY 1. *Consider an observation operator $\mathfrak{X}$ that satisfies the RSC condition (22) with $\gamma > 0$ and $\tau_n = 0$. Suppose that we solve the convex program (10) with regularization parameters (λ_d, μ_d) such that*

$$\lambda_d \geq 4|||\mathfrak{X}^*(W)|||_{\mathrm{op}} \quad \text{and} \quad \mu_d \geq 4\|\mathfrak{X}^*(W)\|_\infty + \frac{4\gamma\alpha}{\sqrt{d_1 d_2}}. \tag{29}$$

Then there are universal constants c_j such that for any matrix pair $(\Theta^\star, \Gamma^\star)$ with $\|\Theta^\star\|_\infty \leq \frac{\alpha}{\sqrt{d_1 d_2}}$ and for all integers $r = 1, 2, \ldots, \min\{d_1, d_2\}$ and $s = 1, 2, \ldots, (d_1 d_2)$, we have the following bound on $e^2(\widehat{\Theta}, \widehat{\Gamma})$:

$$c_1 \frac{\lambda_d^2}{\gamma^2}\left\{r + \frac{\gamma}{\lambda_d}\sum_{j=r+1}^{d}\sigma_j(\Theta^\star)\right\} + c_2\frac{\mu_d^2}{\gamma^2}\left\{s + \frac{\gamma}{\mu_d}\sum_{(j,k)\notin S}|\Gamma^\star_{jk}|\right\}, \tag{30}$$

where S is an arbitrary subset of matrix indices of cardinality at most s.

REMARKS. This result follows directly by specializing Theorem 1 to the elementwise ℓ_1-norm. As noted in Example 4, for this norm, we have $\kappa_d = \sqrt{d_1 d_2}$, so that the choice (29) satisfies the conditions of Theorem 1. The dual norm is given by the elementwise ℓ_∞-norm $\mathcal{R}^*(\cdot) = \|\cdot\|_\infty$. As observed in Section 3.1, the ℓ_1-norm is decomposable with respect to subspace pairs of the form $(\mathbb{M}(S), \mathbb{M}^\perp(S))$, for an arbitrary subset S of matrix indices. Moreover, for any subset S of cardinality s, we have $\Psi^2(\mathbb{M}(S)) = s$. It is easy to verify that with this choice, we have

$$\Pi_{\mathbb{M}^\perp}(\Gamma^\star) = \sum_{(j,k)\notin S}|\Gamma^\star_{jk}|,$$

from which the claim follows.

It is worth noting inequality (27) corresponds to a family of upper bounds indexed by r and the subset S. Given an integer $s \in \{1, 2, \ldots, (d_1 d_2)\}$, it is natural to let S index the largest s entries of $\Gamma^\star$ (in absolute value). Moreover, the choice of the pair (r, s) can be adapted to the structure of the matrix. For instance, when $\Theta^\star$ is exactly low rank, and $\Gamma^\star$ is exactly sparse, one natural choice is $r = \mathrm{rank}(\Theta^\star)$, and $s = |\mathrm{supp}(\Gamma^\star)|$. With this choice, both the approximation terms vanish, and Corollary 1 guarantees that any solution $(\widehat{\Theta}, \widehat{\Gamma})$ of the convex program (10) satisfies

$$|||\widehat{\Theta} - \Theta^\star|||_{\mathrm{F}}^2 + |||\widehat{\Gamma} - \Gamma^\star|||_{\mathrm{F}}^2 \lesssim \lambda_d^2 r + \mu_d^2 s. \tag{31}$$

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

$$|||\widehat{\Theta} - \Theta^\star|||_{\mathrm{F}}^2 + |||\widehat{\Gamma} - \Gamma^\star|||_{\mathrm{F}}^2 \lesssim \alpha^2 \frac{s}{d_1 d_2}. \tag{32}$$

This guarantee is weaker than the exact recovery results obtained in past work on the noiseless observation model with identity operator [7, 9]; however, these papers imposed incoherence requirements on the low-rank component $\Theta^\star$ that are more restrictive than the conditions of Corollary 1.

Our elementwise ℓ_∞ bound is a weaker condition than incoherence, since it allows for singular vectors to be coherent as long as the associated singular value is not too large. Moreover, the bound (32) is optimal up to constant factors, due to the nonidentifiability of the observation model (1), as shown by the following example for the identity observation operator $\mathfrak{X} = I$.

EXAMPLE 6 (Unimprovability for elementwise sparse model). Consider a given sparsity index $s \in \{1, 2, \ldots, (d_1 d_2)\}$, where we may assume without loss of generality that $s \leq d_2$. We then form the matrix

$$\Theta^\star := \frac{\alpha}{\sqrt{d_1 d_2}} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \underbrace{[1 \;\; 1 \;\; 1 \;\; \cdots \;\; 0 \;\; \cdots \;\; 0]}_{f^T}, \tag{33}$$

where the vector $f \in \mathbb{R}^{d_2}$ has exactly s ones. Note that $\|\Theta^\star\|_\infty = \frac{\alpha}{\sqrt{d_1 d_2}}$ by construction, and moreover $\Theta^\star$ is rank one, and has s nonzero entries. Since up to s entries of the noise matrix $\Gamma^\star$ can be chosen arbitrarily, "nature" can always set $\Gamma^\star = -\Theta^\star$, meaning that we would observe $Y = \Theta^\star + \Gamma^\star = 0$. Consequently, based on observing only Y, the pair $(\Theta^\star, \Gamma^\star)$ is indistinguishable from the all-zero matrices $(0_{d_1 \times d_2}, 0_{d_1 \times d_2})$. This fact can be used to show that no method can have squared Frobenius error lower than $\approx \frac{\alpha^2 s}{d_1 d_2}$; see Section 4 for a precise statement. Therefore, bound (32) cannot be improved without imposing further restrictions on the pair $(\Theta^\star, \Gamma^\star)$. We note that the singular vector incoherence conditions, as imposed in past work [7, 9, 11] to guarantee exact recovery, would exclude the matrix (33), since its left singular vector is the unit vector $e_1 \in \mathbb{R}^{d_1}$.

3.4.1. *Results for stochastic noise matrices.* Our discussion thus far has applied to general observation operators $\mathfrak{X}$, and general noise matrices W. More concrete results can be obtained by assuming particular forms of $\mathfrak{X}$, and that the noise matrix W is stochastic. Our first stochastic result applies to the identity operator $\mathfrak{X} = I$ and a noise matrix W generated with i.i.d. $N(0, \nu^2/(d_1 d_2))$ entries.[7]

[7] To be clear, we state our results in terms of the noise scaling $\nu^2/(d_1 d_2)$ since it corresponds to a model with constant signal-to-noise ratio when the Frobenius norms of $\Theta^\star$ and $\Gamma^\star$ remain bounded,

COROLLARY 2. *Suppose* $\mathfrak{X} = I$, *the matrix* $\Theta^\star$ *has rank at most* r *and satisfies* $\|\Theta^\star\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}$, *and* $\Gamma^\star$ *has at most* s *nonzero 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* (10) *with regularization parameters*

$$\lambda_d = \frac{8\nu}{\sqrt{d_1}} + \frac{8\nu}{\sqrt{d_2}} \quad \text{and} \quad \mu_d = 16\nu\sqrt{\frac{\log(d_1 d_2)}{d_1 d_2}} + \frac{4\alpha}{\sqrt{d_1 d_2}}, \tag{34}$$

then with probability greater than $1 - \exp(-2\log(d_1 d_2))$, *any optimal solution* $(\widehat{\Theta}, \widehat{\Gamma})$ *satisfies*

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) \le \underbrace{c_1 \nu^2 \left(\frac{r(d_1 + d_2)}{d_1 d_2} \right)}_{\mathcal{K}^\star_\Theta} + \underbrace{c_1 \nu^2 \left(\frac{s \log(d_1 d_2)}{d_1 d_2} \right) + c_1 \frac{\alpha^2 s}{d_1 d_2}}_{\mathcal{K}^\star_\Gamma}. \tag{35}$$

REMARKS. In the statement of this corollary, the settings of λ_d and μ_d are based on upper bounding $\|W\|_\infty$ and $|||W|||_{\mathrm{op}}$, using large deviation bounds and some nonasymptotic random matrix theory. With a slightly modified argument, bound (35) can be sharpened slightly by reducing the logarithmic term to $\log(\frac{d_1 d_2}{s})$. As shown in Theorem 2 to follow, this sharpened 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.

It is also worth observing that both terms in the bound (35) have intuitive interpretations. Considering first the term $\mathcal{K}^\star_\Theta$, we note that the numerator term $r(d_1 + d_2)$ is of the order of the number of free parameters in a rank r matrix of dimensions $d_1 \times d_2$. The multiplicative factor $\frac{\nu^2}{d_1 d_2}$ corresponds to the noise variance in the problem. On the other hand, the term $\mathcal{K}^\star_\Gamma$ measures the complexity of estimating s nonzero entries in a $d_1 \times d_2$ matrix. Note that there are $\binom{d_1 d_2}{s}$ possible subsets of size s, and consequently, the numerator includes a term that scales as $\log \binom{d_1 d_2}{s} \approx s \log(d_1 d_2)$. As before, the multiplicative pre-factor $\frac{\nu^2}{d_1 d_2}$ corresponds to the noise variance. Finally, the second term within $\mathcal{K}^\star_\Gamma$—namely the quantity $\frac{\alpha^2 s}{d_1 d_2}$—arises from the nonidentifiability of the model, and as discussed in Example 6, it cannot be avoided without imposing further restrictions on the pair $(\Gamma^\star, \Theta^\star)$.

We now turn to analysis of the sparse factor analysis problem: as previously introduced in Example 1, this involves estimation of a covariance matrix that has a low-rank plus elementwise sparse decomposition. In this case, given n i.i.d. samples from the unknown covariance matrix $\Sigma = \Theta^\star + \Gamma^\star$, the noise matrix

independently of the dimension. The same results would hold if the noise were not rescaled, modulo the appropriate rescalings of the various terms.

$W \in \mathbb{R}^{d\times d}$ is a recentered Wishart noise; see equation (3). We can use tail bounds for its entries and its operator norm in order to specify appropriate choices of the regularization parameters λ_d and μ_d. We summarize our conclusions in the following corollary:

COROLLARY 3. *Consider the factor analysis model with $n \geq d$ samples, and regularization parameters*

$$\lambda_d = 16|||\sqrt{\Sigma}|||_2\sqrt{\frac{d}{n}} \quad \text{and} \quad \mu_d = 32\rho(\Sigma)\sqrt{\frac{\log d}{n}} + \frac{4\alpha}{d}, \tag{36}$$

where $\rho(\Sigma) = \max_j \Sigma_{jj}$. Then with probability greater than $1 - c_2\exp(-c_3 \times \log(d))$, any optimal solution $(\widehat{\Theta}, \widehat{\Gamma})$ satisfies

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) \leq c_1\left\{|||\Sigma|||_2\frac{rd}{n} + \rho(\Sigma)\frac{s\log d}{n}\right\} + c_1\frac{\alpha^2 s}{d^2}.$$

We note that the condition $n \geq d$ is necessary to obtain consistent estimates in factor analysis models, even in the special case with $\Gamma^\star = I_{d\times d}$, where standard PCA is possible; for example, see Johnstone [12]. Again, the terms in the bound have a natural interpretation: since a matrix of rank r in d dimensions has roughly rd degrees of freedom, we expect to see a term of the order $\frac{rd}{n}$. Similarly, since there are $\log\binom{d^2}{s} \approx s\log d$ subsets of size s in a $d \times d$ matrix, we also expect to see a term of the order $\frac{s\log d}{n}$. Moreover, although we have stated our choices of regularization parameter in terms of $|||\Sigma|||_2$ and $\rho(\Sigma)$, these can be replaced by the analogous versions using the sample covariance matrix $\widehat{\Sigma}$. (By the concentration results that we establish, the population and empirical versions do not differ significantly when $n \geq d$.) Last, we note that in recent work, Fan et al. [10] have studied an alternative method for estimating sparse factor models, involving a combination of thresholding and principal components. They provide various error bounds, but under different conditions on the interaction between the sparse and low-rank components.

3.4.2. *Comparison to Hsu et al.* [11]. This recent work focuses on the problem of matrix decomposition with the $\|\cdot\|_1$-norm, and provides results both for the noiseless and noisy setting. All of their work focuses on the case of exactly low rank and exactly sparse matrices, and deals only with the identity observation operator; in contrast, Theorem 1 in this paper provides an upper bound for general matrix pairs and observation operators. Most relevant is comparison of our ℓ_1-results with exact rank-sparsity constraints to their Theorem 3, which provides various error bounds (in nuclear and Frobenius norm) for such models with additive noise. These bounds are obtained using an estimator similar to our program (10), and in parts of their analysis, they enforce bounds on the ℓ_∞-norm of the solution.

There are two major differences between our results, and those of Hsu et al. First of all, their analysis involves three quantities (α, β, γ) that measure singular vector incoherence, and must satisfy a number of inequalities. In contrast, our analysis is based only on a single condition: the "spikiness" condition on the low-rank component $\Theta^\star$. As we have seen, this constraint is weaker than singular vector incoherence, and consequently, unlike the result of Hsu et al., we do not provide exact recovery guarantees for the noiseless setting. However, as our analysis shows, a very simple spikiness condition suffices for the approximate recovery guarantees that are of interest for noisy observation models. Given these differing assumptions, the underlying proof techniques are quite distinct, with our methods leveraging the notion of restricted strong convexity introduced by Negahban et al. [14].

The second (and perhaps most significant) difference is in the sharpness of the results for the noisy setting, and the permissible scalings of the rank-sparsity pair (r, s). As demonstrated in Section 4, the rates that we establish for the noisy Gaussian model (Corollary 2) are minimax-optimal up to constant factors, and involve terms *additive* in r and s. In contrast, the upper bounds in Theorem 3 of Hsu et al. involve the *product* rs, and require

$$rs \precsim \frac{d_1 d_2}{\log(d_1)\log(d_2)}. \tag{37}$$

This bound precludes many scalings that are of interest. For instance, if the sparse component $\Gamma^\star$ has a nearly constant fraction of nonzeros [say $s \asymp \frac{d_1 d_2}{\log(d_1)\log(d_2)}$ for concreteness], then bound (37) restricts to $\Theta^\star$ to have constant rank. In contrast, our analysis allows for high-dimensional scaling of both the rank r and sparsity s simultaneously; as can be seen by inspection of Corollary 2, our Frobenius norm error goes to zero under the scalings $s \asymp \frac{d_1 d_2}{\log(d_1)\log(d_2)}$ and $r \asymp \frac{d_2}{\log(d_2)}$.

3.4.3. *Results for multi-task regression.* Let us now extend our results to the setting of multi-task regression, as introduced in Example 2. The observation model is of the form $Y = XB^* + W$, where $X \in \mathbb{R}^{n \times d_1}$ is a known design matrix, and we observe the matrix $Y = \mathbb{R}^{n \times d_2}$. Our goal is to estimate the the regression matrix $B^* \in \mathbb{R}^{d_1 \times d_2}$, which is assumed to have a decomposition of the form $B^* = \Theta^\star + \Gamma^\star$, where $\Theta^\star$ models the shared characteristics between each of the tasks, and the matrix $\Gamma^\star$ models perturbations away from the shared structure. If we take $\Gamma^\star$ to be a sparse matrix, an appropriate choice of regularizer $\mathcal{R}$ is the elementwise ℓ_1-norm, as in Corollary 2. We use $\sigma_{\min}$ and $\sigma_{\max}$ to denote the minimum and maximum singular values (respectively) of the rescaled design matrix $X/\sqrt{n}$; in addition, we assume that X is invertible so that $\sigma_{\min} > 0$, and that its columns are uniformly bounded in ℓ_2-norm (i.e., $\max_{j=1,\ldots,d_1} \|X_j\|_2 \leq \kappa_{\max}\sqrt{n}$). We note that these assumptions are satisfied for many common examples of random design.

COROLLARY 4. *Suppose that the matrix $\Theta^\star$ has rank at most r and satisfies $\|\Theta^\star\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}$, and the matrix $\Gamma^\star$ has at most s nonzero entries. If the entries of W are i.i.d. $N(0, \nu^2)$, and we solve the convex program (10) with regularization parameters*

$$\begin{aligned} \lambda_d &= 8\nu\sigma_{\max}\sqrt{n}(\sqrt{d_1}+\sqrt{d_2}) \quad \text{and} \\ \mu_d &= 16\nu\kappa_{\max}\sqrt{n\log(d_1 d_2)} + \frac{4\alpha\sigma_{\min}\sqrt{n}}{\sqrt{d_1 d_2}}, \end{aligned} \tag{38}$$

then with probability greater than $1 - \exp(-2\log(d_1 d_2))$, the Frobenius error $e^2(\widehat{\Theta}, \widehat{\Gamma})$ of any optimal solution $(\widehat{\Theta}, \widehat{\Gamma})$ is upper bounded by

$$c_1 \underbrace{\frac{\nu^2\sigma_{\max}^2}{\sigma_{\min}^4}\left(\frac{r(d_1+d_2)}{n}\right)}_{\mathcal{K}_\Theta^\star} + c_2 \underbrace{\left[\frac{\nu^2\kappa_{\max}^2}{\sigma_{\min}^4}\left(\frac{s\log(d_1 d_2)}{n}\right) + \frac{\alpha^2 s}{d_1 d_2}\right]}_{\mathcal{K}_\Gamma^\star}. \tag{39}$$

REMARKS. We see that the results presented above are analogous to those presented in Corollary 2. However, in this setting, we leverage large deviations results in order to find bounds on $\|\mathfrak{X}^*(W)\|_\infty$ and $|||\mathfrak{X}^*(W)|||_{\mathrm{op}}$ that hold with high probability, given our observation model.

3.5. *An alternative two-step method.* As suggested by one reviewer, it is possible that a simpler two-step method—namely, based on first thresholding the entries of the observation matrix Y, and then performing a low-rank approximation—might achieve similar rates to the more complex convex relaxation (10). In this section, we provide a detailed analysis of one version of such a procedure in the case of nuclear norm combined with ℓ_1-regularization. We prove that in the special case of $\mathfrak{X} = I$, this procedure can attain the same form of error bounds, with possibly different constants. However, we also give an example to show that the two-step method will not necessarily perform well for general observation operators $\mathfrak{X}$.

In detail, let us consider the following two-step estimator:

(a) Estimate the sparse component $\Gamma^\star$ by solving

$$\widehat{\Gamma} \in \operatorname*{argmin}_{\Gamma \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2}|||Y - \Gamma|||_F^2 + \mu_d \|\Gamma\|_1 \right\}. \tag{40}$$

As is well-known, this convex program has an explicit solution based on soft-thresholding the entries of Y.

(b) Given $\widehat{\Gamma}$, estimate the low-rank component $\Theta^\star$ by computing

$$\widehat{\Theta} \in \operatorname*{argmin}_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2}|||Y - \Theta - \widehat{\Gamma}|||_F^2 + \lambda_d |||\Theta|||_{\mathrm{N}} \right\}. \tag{41}$$

Interestingly, note that this method can be understood as the first two steps of a blockwise co-ordinate descent method for solving the convex program (10). In step (a), we fix the low-rank component, and minimize as a function of the sparse component. In step (b), we fix the sparse component, and then minimize as a function of the low-rank component. The following result that these two steps of co-ordinate descent achieve the same rates (up to constant factors) as solving the full convex program (10):

PROPOSITION 1. Given observations Y from the model $Y = \Theta^\star + \Gamma^\star + W$ with $\|\Theta^\star\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}$, consider the two-step procedure (40) and (41) with regularization parameters (λ_d, μ_d) such that

$$\lambda_d \ge 4|||W|||_{\mathrm{op}} \quad \text{and} \quad \mu_d \ge 4\|W\|_\infty + \frac{4\alpha}{\sqrt{d_1 d_2}}. \tag{42}$$

Then the error bound (30) from Corollary 1 holds with $\gamma = 1$.

Consequently, in the special case that $\mathfrak{X} = I$, just two steps of co-ordinate descent are sufficient to obtain an optimal estimator.

On the other hand, the simple two-stage method will not work for general observation operators $\mathfrak{X}$. As shown in the proof of Proposition 1, the two-step method relies critically on having the quantity $\|\mathfrak{X}(\Theta^\star + W)\|_\infty$ be upper bounded (up to constant factors) by $\max\{\|\Theta^\star\|_\infty, \|W\|_\infty\}$. By triangle inequality, this condition holds trivially when $\mathfrak{X} = I$, but can be violated by other choices of the observation operator, as illustrated below.

EXAMPLE 7 (Failure of two-step method). Recall the multi-task observation model first introduced in Example 2. In Corollary 4, we showed that the general estimator (10) will recover good estimates under certain assumptions on the observation matrix. In this example, we provide an instance for which the assumptions of Corollary 4 are satisfied, but on the other hand, the two-step method will not return a good estimate.

More specifically, let us consider the multivariate regression observation model $Y = X(\Theta^\star + \Gamma^\star) + W$, in which $Y \in \mathbb{R}^{d\times d}$. Suppose that the observation matrix $X \in \mathbb{R}^{d\times d}$ takes the form

$$X := I_{d\times d} + \frac{1}{\sqrt{d}} e_1 \vec{1}^T,$$

where $e_1 \in \mathbb{R}^d$ is the standard basis vector with a 1 in the first component, and $\vec{1}$ denotes the vector of all ones. Suppose that $\Theta^\star = \frac{1}{d}\vec{1}\vec{1}^T$, which is rank one, and satisfies $\|\Theta^\star\|_\infty = \frac{1}{d}$.

We now verify that the conditions of Corollary 4 are satisfied. By construction we have $\sigma_{\min}(X) = 1$ and $\sigma_{\max}(X) \le 2$. Moreover, letting X_j denote the jth column of X, we have $\max_{j=1,\ldots,d} \|X_j\|_2 \le 2$. Consequently, if we consider rescaled

observations with noise variance ν^2/d, the conditions of Corollary 4 are all satisfied with constants (independent of dimension), so that the M-estimator (10) will have good performance.

In comparison, for any zero-mean noise matrix W, we have in expectation

$$\mathbb{E}[\|X(\Theta^\star + W)\|_\infty] \overset{\text{(i)}}{\geq} \|X(\Theta^\star + \mathbb{E}[W])\|_\infty = \|X(\Theta^\star)\|_\infty \overset{\text{(ii)}}{\geq} \sqrt{d}\|\Theta^\star\|_\infty,$$

where step (i) exploits Jensen's inequality, and step (ii) uses the fact that

$$\|X(\Theta^\star)\|_\infty = 1/d + 1/\sqrt{d} = (1+\sqrt{d})\|\Theta^\star\|_\infty.$$

For any noise matrix W with reasonable tail behavior, the random variable $\|X(\Theta^\star + W)\|_\infty$ will concentrate around its expectation, showing that $\|X(\Theta^\star + W)\|_\infty$ will be larger than $\|\Theta^\star\|_\infty$ by an order of magnitude (factor of $\sqrt{d}$). Consequently, the two-step method will have much larger error in this case.

3.6. *Results for $\|\cdot\|_{2,1}$ regularization.* Let us return again to the general Theorem 1, and illustrate some more of its consequences in application to the columnwise $(2,1)$-norm previously defined in Example 5, and methods based on solving the convex program (14). As before, specializing Theorem 1 to this decomposable regularizer yields a number of guarantees. In order to keep our presentation relatively brief, we focus here on the case of the identity observation operator $\mathfrak{X} = I$.

COROLLARY 5. *Suppose that we solve the convex program* (14) *with regularization parameters* (λ_d, μ_d) *such that*

$$\lambda_d \geq 4|||W|||_{\text{op}} \quad \text{and} \quad \mu_d \geq 4\|W\|_{2,\infty} + \frac{4\alpha}{\sqrt{d_2}}. \tag{43}$$

Then there is a universal constant c_1 such that for any matrix pair $(\Theta^\star, \Gamma^\star)$ with $\|\Theta^\star\|_{2,\infty} \leq \frac{\alpha}{\sqrt{d_2}}$ and for all integers $r = 1, 2, \ldots, d$ and $s = 1, 2, \ldots, d_2$, we have the following bound on $e^2(\widehat{\Theta}, \widehat{\Gamma})$:

$$c_1\lambda_d^2\left\{r + \frac{1}{\lambda_d}\sum_{j=r+1}^{d}\sigma_j(\Theta^\star)\right\} + c_1\mu_d^2\left\{s + \frac{1}{\mu_d}\sum_{k\notin C}\|\Gamma_k^\star\|_2\right\}, \tag{44}$$

where $C \subseteq \{1, 2, \ldots, d_2\}$ is an arbitrary subset of column indices of cardinality at most s.

REMARKS. This result follows directly by specializing Theorem 1 to the columnwise $(2,1)$-norm and identity observation model, discussed in Example 5. Its dual norm is the $(2,\infty)$-norm, and we have $\kappa_d = \sqrt{d_2}$. As discussed in Section 3.1, the $(2,1)$-norm is decomposable with respect to subspaces of the type $\mathbb{M}(C)$, as defined in equation (17), where C is an arbitrary subset of columns. For any such subset C of cardinality s, it can be calculated that $\Psi^2(\mathbb{M}(C)) = s$,

and $\|\Pi_{\mathbb{M}^\perp}(\Gamma^\star)\|_{2,1} = \sum_{k \notin C} \|\Gamma^\star_k\|_2$. Consequently, bound (44) follows from Theorem 1.

As before, if we assume that $\Theta^\star$ has exactly rank r and $\Gamma^\star$ has at most s nonzero columns, then both approximation error terms in bound (44) vanish, and we recover an upper bound of the form $|||\widehat{\Theta} - \Theta^\star|||_{\mathrm{F}}^2 + |||\widehat{\Gamma} - \Gamma^\star|||_{\mathrm{F}}^2 \lesssim \lambda_d^2 r + \mu_d^2 s$. If we further specialize to the case of exact observations ($W = 0$), then Corollary 5 guarantees that

$$|||\widehat{\Theta} - \Theta^\star|||_{\mathrm{F}}^2 + |||\widehat{\Gamma} - \Gamma^\star|||_{\mathrm{F}}^2 \lesssim \alpha^2 \frac{s}{d_2}.$$

The following example shows, that given our conditions, even in the noiseless setting, no method can do better.

EXAMPLE 8 (Unimprovability for columnwise sparse model). In order to demonstrate that the term $\alpha^2 s/d_2$ is unavoidable, it suffices to consider a slight modification of Example 6. In particular, let us define the matrix

$$\Theta^\star := \frac{\alpha}{\sqrt{d_1 d_2}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \underbrace{[1 \;\; 1 \;\; 1 \;\; \cdots \;\; 0 \;\; \cdots \;\; 0]}_{f^T}, \tag{45}$$

where again the vector $f \in \mathbb{R}^{d_2}$ has s nonzeros. Note that the matrix $\Theta^\star$ is rank one, has s nonzero columns, and moreover $\|\Theta^\star\|_{2,\infty} = \frac{\alpha}{\sqrt{d_2}}$. Consequently, the matrix $\Theta^\star$ is covered by Corollary 5. Since s columns of the matrix $\Gamma^\star$ can be chosen in an arbitrary manner, it is possible that $\Gamma^\star = -\Theta^\star$, in which case the observation matrix $Y = 0$. This fact can be exploited to show that *no method* can achieve squared Frobenius error much smaller than $\approx \frac{\alpha^2 s}{d_2}$; see Section 4 for the precise statement. Finally, we note that it is difficult to compare directly to the results of Xu et al. [21], since their results do not guarantee exact recovery of the pair $(\Theta^\star, \Gamma^\star)$.

As with the case of elementwise ℓ_1-norm, more concrete results can be obtained when the noise matrix W is stochastic.

COROLLARY 6. *Suppose $\Theta^\star$ has rank at most r and satisfies $\|\Theta^\star\|_{2,\infty} \leq \frac{\alpha}{\sqrt{d_2}}$, and $\Gamma^\star$ has at most s nonzero 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 (14) with regularization parameters $\lambda_d = \frac{8\nu}{\sqrt{d_1}} + \frac{8\nu}{\sqrt{d_2}}$ and*

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

then with probability greater than $1 - \exp(-2\log(d_2))$, *any optimal solution* $(\widehat{\Theta}, \widehat{\Gamma})$ *satisfies*

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) \leq c_1 \underbrace{\nu^2 \frac{r(d_1+d_2)}{d_1 d_2}}_{\mathcal{K}^\star_\Theta} + \underbrace{\nu^2 \left\{ \frac{sd_1}{d_1 d_2} + \frac{s\log d_2}{d_1 d_2} \right\}}_{\mathcal{K}^\star_\Gamma} + c_2 \frac{\alpha^2 s}{d_2}. \tag{46}$$

REMARKS. Note that the setting of λ_d is the same as in Corollary 2, whereas the parameter μ_d is chosen based on upper bounding $\|W\|_{2,\infty}$, corresponding to the dual norm of the columnwise $(2,1)$-norm. With a slightly modified argument, bound (46) can be sharpened slightly by reducing the logarithmic term to $\log(\frac{d_2}{s})$. As shown in Theorem 2, this sharpened bound is minimax-optimal.

As with Corollary 2, both terms in the bound (46) are readily interpreted. The term $\mathcal{K}^\star_\Theta$ has the same interpretation, as a combination of the number of degrees of freedom in a rank r matrix [i.e., of the order $r(d_1+d_2)$], scaled by the noise variance $\frac{\nu^2}{d_1 d_2}$. The second term $\mathcal{K}^\star_\Gamma$ has a somewhat more subtle interpretation. The problem of estimating s nonzero columns embedded within a $d_1 \times d_2$ matrix can be split into two sub-problems: first, the problem of estimating the sd_1 nonzero parameters (in Frobenius norm), and second, the problem of column subset selection—that is, determining the location of the s nonzero parameters. The estimation sub-problem yields the term $\frac{\nu^2 sd_1}{d_1 d_2}$, whereas the column subset selection sub-problem incurs a penalty involving $\log\binom{d_2}{s} \approx s\log d_2$, multiplied by the usual noise variance. The final term $\alpha^2 s/d_2$ arises from the nonidentifiability of the model. As discussed in Example 8, it is unavoidable without further restrictions.

We now turn to some consequences for the problem of robust covariance estimation formulated in Example 3. As seen from equation (4), the disturbance matrix in this setting can be written as a sum $(\Gamma^\star)^T + \Gamma^\star$, where $\Gamma^\star$ is a columnwise sparse matrix. Consequently, we can use a variant of the estimator (14), in which the loss function is given by $|||Y - \{\Theta^\star + (\Gamma^\star)^T + \Gamma^\star\}|||_F^2$. The following result gives the consequences of Theorem 1 in this setting:

COROLLARY 7. *Consider the problem of robust covariance estimation with* $n \geq d$ *samples, based on a matrix* $\Theta^\star$ *with rank at most* r *that satisfies* $\|\Theta^\star\|_{2,\infty} \leq \frac{\alpha}{\sqrt{d}}$, *and a corrupting matrix* $\Gamma^\star$ *with at most* s *rows and columns corrupted. If we solve SDP* (14) *with regularization parameters*

$$\lambda_d^2 = 8|||\Theta^\star|||_{\mathrm{op}}^2 \frac{r}{n} \quad \text{and} \quad \mu_d^2 = 8|||\Theta^\star|||_{\mathrm{op}}^2 \frac{r}{n} + \frac{16\alpha^2}{d}, \tag{47}$$

then with probability greater than $1 - c_2\exp(-c_3\log(d))$, *any optimal solution* $(\widehat{\Theta}, \widehat{\Gamma})$ *satisfies*

$$e^2(\widehat{\Theta}, \widehat{\Gamma}) \leq c_1 |||\Theta^\star|||_{\mathrm{op}}^2 \left\{ \frac{r^2}{n} + \frac{sr}{n} \right\} + c_2 \frac{\alpha^2 s}{d}.$$

Some comments about this result: with the motivation of being concrete, we have given an explicit choice (47) of the regularization parameters, involving the operator norm $|||\Theta^\star|||_{\mathrm{op}}$, but any upper bound would suffice. As with the noise variance in Corollary 6, a typical strategy would choose this pre-factor by cross-validation.

4. Lower bounds. For the case of i.i.d. Gaussian noise matrices, Corollaries 2 and 6 provide results of an achievable nature, namely in guaranteeing that our estimators achieve certain Frobenius norm errors. In this section, we turn to the complementary question: what are the fundamental (algorithmic-independent) limits of accuracy in noisy matrix decomposition? One way in which to address such a question is by analyzing statistical minimax rates.

More formally, given some family $\mathcal{F}$ of matrices, the associated minimax error is given by

$$\mathfrak{M}(\mathcal{F}) := \inf_{(\widetilde{\Theta},\widetilde{\Gamma})} \sup_{(\Theta^\star,\Gamma^\star)} \mathbb{E}[|||\widetilde{\Theta} - \Theta^\star|||_{\mathrm{F}}^2 + |||\widetilde{\Gamma} - \Gamma^\star|||_{\mathrm{F}}^2], \tag{48}$$

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).

Given a matrix $\Gamma^\star$, we define its support set $\mathrm{supp}(\Gamma^\star) := \{(j,k) \mid \Gamma^\star_{jk} \neq 0\}$, as well as its column support $\mathrm{colsupp}(\Gamma^\star) := \{k \mid \Gamma^\star_k \neq 0\}$, where $\Gamma^\star_k$ denotes the kth column. Using this notation, our interest centers on the following two matrix families:

$$\mathcal{F}_{\mathrm{sp}}(r,s,\alpha) := \Big\{(\Theta^\star,\Gamma^\star) \mid \mathrm{rank}(\Theta^\star) \le r, |\mathrm{supp}(\Gamma^\star)| \le s, \|\Theta^\star\|_\infty \le \frac{\alpha}{\sqrt{d_1 d_2}}\Big\}, \tag{49a}$$

and

$$\mathcal{F}_{\mathrm{col}}(r,s,\alpha) := \Big\{(\Theta^\star,\Gamma^\star) \mid \mathrm{rank}(\Theta^\star) \le r, |\mathrm{colsupp}(\Gamma^\star)| \le s, \|\Theta^\star\|_{2,\infty} \le \frac{\alpha}{\sqrt{d_2}}\Big\}. \tag{49b}$$

By construction, Corollaries 2 and 6 apply to the families $\mathcal{F}_{\mathrm{sp}}$ and $\mathcal{F}_{\mathrm{col}}$, respectively.

The following theorem establishes lower bounds on the minimax risks (in squared Frobenius norm) over these two families for the identity observation operator:

THEOREM 2. *Consider the linear observation model* (1) *with identity observation operator*: $\mathfrak{X}(\Theta + \Gamma) = \Theta + \Gamma$. *There is a universal constant* $c_0 > 0$ *such that for all* $\alpha \geq 32\sqrt{\log(d_1 d_2)}$, *we have*

$$\mathfrak{M}(\mathcal{F}_{\mathrm{sp}}(r, s, \alpha)) \geq c_0 \nu^2 \left\{ \frac{r(d_1 + d_2)}{d_1 d_2} + s \log\left(\frac{d_1 d_2 - s}{s/2} \right) / (d_1 d_2) \right\} + c_0 \frac{\alpha^2 s}{d_1 d_2}, \tag{50}$$

and

$$\mathfrak{M}(\mathcal{F}_{\mathrm{col}}(r, s, \alpha)) \geq c_0 \nu^2 \left(\frac{r(d_1 + d_2)}{d_1 d_2} + \frac{s}{d_2} + s \log\left(\frac{d_2 - s}{s/2} \right) / (d_1 d_2) \right) + c_0 \frac{\alpha^2 s}{d_2}. \tag{51}$$

Note the agreement with the achievable rates guaranteed in Corollaries 2 and 6, respectively. (As discussed in the remarks following these corollaries, the sharpened forms of the logarithmic factors follow by a more careful analysis.) Theorem 2 shows that in terms of squared Frobenius error, the convex relaxations (10) and (14) are minimax optimal up to constant factors.

In addition, it is worth observing that although Theorem 2 is stated in the context of additive Gaussian noise, it also shows that the radius of nonidentifiability (involving the parameter α) is a fundamental limit. In particular, by setting the noise variance to zero, we see that under our milder conditions, even in the noiseless setting, no algorithm can estimate to greater accuracy than $c_0 \frac{\alpha^2 s}{d_1 d_2}$, or the analogous quantity for column-sparse matrices.

5. Discussion. In this paper, we analyzed a class of convex relaxations for solving a general class of matrix decomposition problems, in which the goal is to recover a pair of matrices, based on observing a noisy contaminated version of their sum. Since the problem is ill-posed in general, it is essential to impose structure, and this paper focuses on the setting in which one matrix is approximately low-rank, and the second has a complementary form of low-dimensional structure enforced by a decomposable regularizer. Particular cases include matrices that are elementwise sparse, or columnwise sparse, and the associated matrix decomposition problems have various applications, including robust PCA, robustness in collaborative filtering, and model selection in Gauss–Markov random fields. We provided a general nonasymptotic bound on the Frobenius norm error of a convex relaxation based on regularizing norm the least-squares loss with a combination of the nuclear norm with a decomposable regularizer. When specialized to the case of elementwise and columnwise sparsity, these estimators yield rates that are minimax-optimal up to constant factors.

Various extensions of this work are possible. We have not discussed here how our estimator would behave under a partial observation model, in which only a

fraction of the entries are observed. This problem is very closely related to matrix completion, a problem for which recent work by Negahban and Wainwright [16] shows that a form of restricted strong convexity holds with high probability. This property could be adapted to the current setting, and would allow for proving Frobenius norm error bounds on the low rank component. Finally, although this paper has focused on the case in which the first matrix component is approximately low rank, much of our theory could be applied to a more general class of matrix decomposition problems, in which the first component is penalized by a decomposable regularizer that is "complementary" to the second matrix component. It remains to explore the properties and applications of these general matrix decompositions.

Acknowledgments. All three authors would like to acknowledge the Banff International Research Station (BIRS) in Banff, Canada for hospitality and work facilities that stimulated and supported this collaboration. The authors also thank anonymous reviewers and editors for thoughtful comments and suggestions which significantly clarified and improved the manuscript.

SUPPLEMENTARY MATERIAL

Simulations and proofs (DOI: 10.1214/12-AOS1000SUPP; .pdf). This supplementary material contains numerical simulations that demonstrate excellent agreement between the theoretical predictions and the practical behavior of our estimators. We also provide proofs for our upper and lower bounds, including slightly sharpened versions of Corollaries 2 and 6.

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Supplement to "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions." DOI:10.1214/12-AOS1000SUPP.

[2] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

[3] ANDO, R. K. and ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6** 1817–1853. MR2249873

[4] BLITZER, J., FOSTER, D. P. and KAKADE, S. M. (2009). Zero-shot domain adaptation: A multi-view approach. Technical report, Toyota Technological Institute at Chicago.

[5] BLITZER, J., MCDONALD, R. and PEREIRA, F. (2006). Domain adaptation with structural correspondence learning. In *EMNLP Conference, Sydney, Australia*.

[6] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

[7] CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000

[8] CHANDRASEKARAN, V., PARILLO, P. A. and WILLSKY, A. S. (2010). Latent variable graphical model selection via convex optimization. Technical report, Massachusetts Institute of Technology.

[9] CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

[10] FAN, J., LIAO, Y. and MINCHEVA, M. (2012). Large covariance estimation by thresholding principal orthogonal components. Technical report, Princeton Univ. Available at arXiv:1201.0175v1.

[11] HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory* **57** 7221–7234. MR2883652

[12] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[13] MCCOY, M. and TROPP, J. A. (2011). Two proposals for robust PCA using semidefinite programming. *Electron. J. Stat.* **5** 1123–1160. MR2836771

[14] NEGAHBAN, S., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS Conference*, Vancouver, Canada, December 2009. Full length version available at arXiv:1010.2731v1. *Statist. Sci.* To appear.

[15] NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

[16] NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697.

[17] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

[18] RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543

[19] ROCKAFELLAR, R. T. (1970). *Convex Analysis. Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. MR0274683

[20] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

[21] XU, H., CARAMANIS, C. and SANGHAVI, S. (2010). Robust PCA via outlier pursuit. Technical report, Univ. Texas, Austin. Available at arXiv:1010.4237.

[22] YUAN, M., EKICI, A., LU, Z. and MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 329–346. MR2323756

A. AGARWAL
DEPARTMENT OF EECS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: alekh@eecs.berkeley.edu

S. NEGAHBAN
DEPARTMENT OF EECS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
32 VASSAR STREET
CAMBIRDGE, MASSACHUSETTS 02139
USA
E-MAIL: sahandn@mit.edu

M. J. WAINWRIGHT
DEPARTMENT OF EECS AND STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@stat.berkeley.edu