The Annals of Statistics
2011, Vol. 39, No. 2, 1069–1097
DOI: 10.1214/10-AOS850

ESTIMATION OF (NEAR) LOW-RANK MATRICES WITH NOISE AND HIGH-DIMENSIONAL SCALING

BY SAHAND NEGAHBAN AND MARTIN J. WAINWRIGHT[1,2]

University of California, Berkeley

We study an instance of high-dimensional inference in which the goal is to estimate a matrix $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$ on the basis of N noisy observations. The unknown matrix Θ^* is assumed to be either exactly low rank, or "near" low-rank, meaning that it can be well-approximated by a matrix with low rank. We consider a standard M-estimator based on regularization by the nuclear or trace norm over matrices, and analyze its performance under high-dimensional scaling. We define the notion of restricted strong convexity (RSC) for the loss function, and use it to derive nonasymptotic bounds on the Frobenius norm error that hold for a general class of noisy observation models, and apply to both exactly low-rank and approximately low rank matrices. We then illustrate consequences of this general theory for a number of specific matrix models, including low-rank multivariate or multi-task regression, system identification in vector autoregressive processes and recovery of low-rank matrices from random projections. These results involve nonasymptotic random matrix theory to establish that the RSC condition holds, and to determine an appropriate choice of regularization parameter. Simulation results show excellent agreement with the high-dimensional scaling of the error predicted by our theory.

1. Introduction. High-dimensional inference refers to instances of statistical estimation in which the ambient dimension of the data is comparable to (or possibly larger than) the sample size. Problems with a high-dimensional character arise in a variety of applications in science and engineering, including analysis of gene array data, medical imaging, remote sensing and astronomical data analysis. In settings where the number of parameters may be large relative to the sample size, the utility of classical (fixed dimension) results is questionable, and accordingly, a line of on-going statistical research seeks to obtain results that hold under high-dimensional scaling, meaning that both the problem size and sample size (as well as other problem parameters) may tend to infinity simultaneously. It is usually impossible to obtain consistent procedures in such settings without imposing some sort of additional constraints. Accordingly, there are now various lines of

Received January 2010; revised August 2010.
[1]Supported in part by a Sloan Foundation Fellowship, Grant AFOSR-09NL184.
[2]Supported by NSF Grants CDI-0941742 and DMS-09-07632.
MSC2010 subject classifications. Primary 62F30; secondary 62H12.
Key words and phrases. High-dimensional inference, rank constraints, nuclear norm, trace norm, M-estimators, random matrix theory.

work on high-dimensional inference based on imposing different types of structural constraints. A substantial body of past work has focused on models with sparsity constraints, including the problem of sparse linear regression [10, 16, 18, 36, 50], banded or sparse covariance matrices [7, 8, 19], sparse inverse covariance matrices [24, 44, 48, 55], sparse eigenstructure [2, 30, 42], and sparse regression matrices [28, 34, 41, 54]. A theme common to much of this work is the use of the ℓ_1-penalty as a surrogate function to enforce the sparsity constraint. A parallel line of work has focused on the use of concave penalties to achieve gains in model selection and sparsity recovery [20, 21].

In this paper, we focus on the problem of high-dimensional inference in the setting of matrix estimation. In contrast to past work, our interest in this paper is the problem of estimating a matrix $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$ that is either *exactly low rank*, meaning that it has at most $r \ll \min\{m_1, m_2\}$ nonzero singular values, or more generally is *near low-rank*, meaning that it can be well-approximated by a matrix of low rank. As we discuss at more length in the sequel, such exact or approximate low-rank conditions are appropriate for many applications, including multivariate or multi-task forms of regression, system identification for autoregressive processes, collaborative filtering, and matrix recovery from random projections. Analogous to the use of an ℓ_1-regularizer for enforcing sparsity, we consider the use of the nuclear norm (also known as the trace norm) for enforcing a rank constraint in the matrix setting. By definition, the nuclear norm is the sum of the singular values of a matrix, and so encourages sparsity in the vector of singular values, or equivalently for the matrix to be low-rank. The problem of low-rank matrix approximation and the use of nuclear norm regularization have been studied by various researchers. In her Ph.D. thesis, Fazel [22] discusses the use of nuclear norm as a heuristic for restricting the rank of a matrix, showing that in practice it is often able to yield low-rank solutions. Other researchers have provided theoretical guarantees on the performance of nuclear norm and related methods for low-rank matrix approximation. Srebro, Rennie and Jaakkola [49] proposed nuclear norm regularization for the collaborative filtering problem, and established risk consistency under certain settings. Recht, Fazel and Parrilo [45] provided sufficient conditions for exact recovery using the nuclear norm heuristic when observing random projections of a low-rank matrix, a set-up analogous to the compressed sensing model in sparse linear regression [14, 18]. Other researchers have studied a version of matrix completion in which a subset of entries are revealed, and the goal is to obtain perfect reconstruction either via the nuclear norm heuristic [15] or by other SVD-based methods [31]. For general observation models, Bach [6] has provided results on the consistency of nuclear norm minimization in noisy settings, but applicable to the classical "fixed p" setting. In addition, Yuan et al. [53] provide nonasymptotic bounds on the operator norm error of the estimate in the multi-task setting, provided that the design matrices are orthogonal. Under the assumption of RIP, Lee and Bresler [32] prove stability properties of least-squares under nuclear norm constraint when a form of restricted isometry property is imposed on the

sampling operator. Liu and Vandenberghe [33] develop an efficient interior-point method for solving nuclear-norm constrained problems and illustrate its usefulness for problems of system identification, an application also considered in this paper. Finally, in work posted shortly after our own, Rohde and Tsybakov [47] and Candes and Plan [13] have studied certain aspects of nuclear norm minimization under high-dimensional scaling. We discuss connections to this concurrent work at more length in Section 3.2 following the statement of our main results.

The goal of this paper is to analyze the nuclear norm relaxation for a general class of noisy observation models and obtain nonasymptotic error bounds on the Frobenius norm that hold under high-dimensional scaling and are applicable to both exactly and approximately low-rank matrices. We begin by presenting a generic observation model and illustrating how it can be specialized to the several cases of interest, including low-rank multivariate regression, estimation of autoregressive processes and random projection (compressed sensing) observations. In particular, this model is specified in terms of an operator $\mathfrak{X}$, which may be deterministic or random depending on the setting, that maps any matrix $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$ to a vector of N noisy observations. We then present a single main theorem (Theorem 1) followed by two corollaries that cover the cases of exact low-rank constraints (Corollary 1) and near low-rank constraints (Corollary 2), respectively. These results demonstrate that high-dimensional error rates are controlled by two key quantities. First, the (random) observation operator $\mathfrak{X}$ is required to satisfy a condition known as *restricted strong convexity* (RSC), introduced in a more general setting by Negahban et al. [37], which ensures that the loss function has sufficient curvature to guarantee consistent recovery of the unknown matrix Θ^*. As we show via various examples, this RSC condition is weaker than the RIP property, which requires that the sampling operator behave very much like an isometry on low-rank matrices. Second, our theory provides insight into the *choice of regularization parameter* that weights the nuclear norm, showing that an appropriate choice is to set it proportionally to the spectral norm of a random matrix defined by the adjoint of observation operator $\mathfrak{X}$, and the observation noise in the problem.

This initial set of results, though appealing in terms of their simple statements and generality, are somewhat abstractly formulated. Our next contribution is to show that by specializing our main result (Theorem 1) to three classes of models, we can obtain some concrete results based on readily interpretable conditions. In particular, Corollary 3 deals with the case of low-rank multivariate regression, relevant for applications in multitask learning. We show that the random operator $\mathfrak{X}$ satisfies the RSC property for a broad class of observation models, and we use random matrix theory to provide an appropriate choice of the regularization parameter. Our next result, Corollary 4, deals with the case of estimating the matrix of parameters specifying a vector autoregressive (VAR) process [4, 35]. The usefulness of the nuclear norm in this context has been demonstrated by Liu and Vandenberghe [33]. Here we also establish that a suitable RSC property holds with high probability for the random operator $\mathfrak{X}$, and also specify a suitable choice

of the regularization parameter. We note that the technical details here are considerably more subtle than the case of low-rank multivariate regression, due to dependencies introduced by the autoregressive sampling scheme. Accordingly, in addition to terms that involve the size, the matrix dimensions and rank, our bounds also depend on the mixing rate of the VAR process. Finally, we turn to the compressed sensing observation model for low-rank matrix recovery, as introduced by Recht and colleagues [45, 46]. In this setting, we again establish that the RSC property holds with high probability, specify a suitable choice of the regularization parameter and thereby obtain a Frobenius error bound for noisy observations (Corollary 5). A technical result that we prove en route—namely, Proposition 1—is of possible independent interest, since it provides a bound on the constrained norm of a random Gaussian operator. In particular, this proposition allows us to obtain a sharp result (Corollary 6) for the problem of recovering a low-rank matrix from perfectly observed random Gaussian projections with a general dependency structure.

The remainder of this paper is organized as follows. Section 2 is devoted to background material, and the set-up of the problem. We present a generic observation model for low-rank matrices, and then illustrate how it captures various cases of interest. We then define the convex program based on nuclear norm regularization that we analyze in this paper. In Section 3, we state our main theoretical results and discuss their consequences for different model classes. Section 4 is devoted to the proofs of our results; in each case, we break down the key steps in a series of lemmas, with more technical details deferred to the appendices. In Section 5, we present the results of various simulations that illustrate excellent agreement between the theoretical bounds and empirical behavior.

NOTATION. For the convenience of the reader, we collect standard pieces of notation here. For a pair of matrices Θ and Γ with commensurate dimensions, we let $\langle\langle \Theta, \Gamma \rangle\rangle = \operatorname{trace}(\Theta^T \Gamma)$ denote the trace inner product on matrix space. For a matrix $\Theta \in \mathbb{R}^{m_1 \times m_2}$, we define $m = \min\{m_1, m_2\}$, and denote its (ordered) singular values by $\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \cdots \geq \sigma_m(\Theta) \geq 0$. We also use the notation $\sigma_{\max}(\Theta) = \sigma_1(\Theta)$ and $\sigma_{\min}(\Theta) = \sigma_m(\Theta)$ to refer to the maximal and minimal singular values, respectively. We use the notation $|||\cdot|||$ for various types of matrix norms based on these singular values, including the *nuclear norm* $|||\Theta|||_1 = \sum_{j=1}^m \sigma_j(\Theta)$, the *spectral or operator norm* $|||\Theta|||_{\mathrm{op}} = \sigma_1(\Theta)$, and the *Frobenius norm* $|||\Theta|||_F = \sqrt{\operatorname{trace}(\Theta^T \Theta)} = \sqrt{\sum_{j=1}^m \sigma_j^2(\Theta)}$. We refer the reader to Horn and Johnson [26, 27] for more background on these matrix norms and their properties.

2. Background and problem set-up. We begin with some background on problems and applications in which rank constraints arise, before describing a generic observation model. We then introduce the semidefinite program (SDP) based on nuclear norm regularization that we study in this paper.

2.1. *Models with rank constraints.* Imposing a rank r constraint on a matrix $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$ is equivalent to requiring the rows (or columns) of Θ^* lie in some r-dimensional subspace of $\mathbb{R}^{m_2}$ (or $\mathbb{R}^{m_1}$, resp.). Such types of rank constraints (or approximate forms thereof) arise in a variety of applications, as we discuss here. In some sense, rank constraints are a generalization of sparsity constraints; rather than assuming that the data is sparse in a known basis, a rank constraint implicitly imposes sparsity but without assuming the basis.

We first consider the problem of multivariate regression, also referred to as multi-task learning in statistical machine learning. The goal of *multivariate regression* is to estimate a prediction function that maps covariates $Z_j \in \mathbb{R}^m$ to multi-dimensional output vectors $Y_j \in \mathbb{R}^{m_1}$. More specifically, let us consider the linear model, specified by a matrix $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$, of the form

$$Y_a = \Theta^* Z_a + W_a \qquad \text{for } a = 1, \ldots, n, \tag{1}$$

where $\{W_a\}_{a=1}^n$ is an i.i.d. sequence of m_1-dimensional zero-mean noise vectors. Given a collection of observations $\{Z_a, Y_a\}_{a=1}^n$ of covariate-output pairs, our goal is to estimate the unknown matrix Θ^*. This type of model has been used in many applications, including analysis of fMRI image data [25], analysis of EEG data decoding [3], neural response modeling [12] and analysis of financial data. This model and closely related ones also arise in the problem of collaborative filtering [49], in which the goal is to predict users' preferences for items (such as movies or music) based on their and other users' ratings of related items. The papers [1, 5] discuss additional instances of low-rank decompositions. In all of these settings, the low-rank condition translates into the existence of a smaller set of "features" that are actually controlling the prediction.

As a second (not unrelated) example, we now consider the problem of system identification in vector autoregressive processes (see [35] for detailed background). A *vector autoregressive* (VAR) process in m-dimensions is a stochastic process $\{Z_t\}_{t=1}^\infty$ specified by an initialization $Z_1 \in \mathbb{R}^m$, followed by the recursion

$$Z_{t+1} = \Theta^* Z_t + W_t \qquad \text{for } t = 1, 2, 3, \ldots. \tag{2}$$

In this recursion, the sequence $\{W_t\}_{t=1}^\infty$ consists of i.i.d. samples of innovations noise. We assume that each vector $W_t \in \mathbb{R}^m$ is zero-mean with covariance matrix $C \succ 0$, so that the process $\{Z_t\}_{t=1}^\infty$ is zero-mean and has a covariance matrix Σ given by the solution of the discrete-time Ricatti equation,

$$\Sigma = \Theta^* \Sigma (\Theta^*)^T + C. \tag{3}$$

The goal of system identification in a VAR process is to estimate the unknown matrix $\Theta^* \in \mathbb{R}^{m \times m}$ on the basis of a sequence of samples $\{Z_t\}_{t=1}^n$. In many application domains, it is natural to expect that the system is controlled primarily by a low-dimensional subset of variables. For instance, models of financial data might have an ambient dimension m of thousands (including stocks, bonds, and other financial instruments), but the behavior of the market might be governed by a much

smaller set of macro-variables (combinations of these financial instruments). Similar statements apply to other types of time series data, including neural data [12, 23], subspace tracking models in signal processing and motion models models in computer vision. While the form of system identification formulated here assumes direct observation of the state variables $\{Z_t\}_{t=1}^n$, it is also possible to tackle the more general problem when only noisy versions are observed (e.g., see Liu and Vandenberghe [33]). An interesting feature of the system identification problem is that the matrix Θ^*, in addition to having low rank, might also be required to satisfy some type of structural constraint (e.g., having a Hankel-type structure), and the estimator that we consider here allows for this possibility.

A third example that we consider in this paper is a *compressed sensing* observation model, in which one observes random projections of the unknown matrix Θ^*. This observation model has been studied extensively in the context of estimating sparse vectors [14, 18], and Recht and colleagues [45, 46] suggested and studied its extension to low-rank matrices. In their set-up, one observes trace inner products of the form $\langle\langle X_i, \Theta^* \rangle\rangle = \operatorname{trace}(X_i^T \Theta^*)$, where $X_i \in \mathbb{R}^{m_1 \times m_2}$ is a random matrix [e.g., filled with standard normal $N(0, 1)$ entries], so that $\langle\langle X_i, \Theta^* \rangle\rangle$ is a standard random projection. In the sequel, we consider this model with a more general family of random projections involving matrices with dependent entries. Like compressed sensing for sparse vectors, applications of this model include computationally efficient updating in large databases (where the matrix Θ^* measures the difference between the data base at two different time instants) and matrix denoising.

2.2. *A generic observation model.* We now introduce a generic observation model that will allow us to deal with these different observation models in an unified manner. For pairs of matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$, recall the Frobenius or trace inner product $\langle\langle A, B \rangle\rangle := \operatorname{trace}(B A^T)$. We then consider a linear observation model of the form

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, N, \tag{4}$$

which is specified by the sequence of observation matrices $\{X_i\}_{i=1}^N$ and observation noise $\{\varepsilon_i\}_{i=1}^N$. This observation model can be written in a more compact manner using operator-theoretic notation. In particular, let us define the observation vector

$$\vec{y} = [\, y_1 \quad \cdots \quad y_N \,]^T \in \mathbb{R}^N$$

with a similar definition for $\vec{\varepsilon} \in \mathbb{R}^N$ in terms of $\{\varepsilon_i\}_{i=1}^N$. We then use the observation matrices $\{X_i\}_{i=1}^N$ to define an operator $\mathfrak{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^N$ via $[\mathfrak{X}(\Theta)]_i = \langle\langle X_i, \Theta \rangle\rangle$. With this notation, the observation model (4) can be re-written as

$$\vec{y} = \mathfrak{X}(\Theta^*) + \vec{\varepsilon}. \tag{5}$$

Let us illustrate the form of the observation model (5) for some of the applications that we considered earlier.

EXAMPLE 1 (Multivariate regression). Recall the observation model (1) for multivariate regression. In this case, we make n observations of vector pairs $(Y_a, Z_a) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$. Accounting for the m_1-dimensional nature of the output, after the model is scalarized, we receive a total of $N = m_1 n$ observations. Let us introduce the quantity $b = 1, \ldots, m_1$ to index the different elements of the output, so that we can write

$$Y_{ab} = \langle\langle e_b Z_a^T, \Theta^* \rangle\rangle + W_{ab} \qquad \text{for } b = 1, 2, \ldots, m_1. \tag{6}$$

By re-indexing this collection of $N = nm_1$ observations via the mapping

$$(a, b) \mapsto i = (a - 1)m_1 + b,$$

we recognize multivariate regression as an instance of the observation model (4) with observation matrix $X_i = e_b Z_a^T$ and scalar observation $y_i = Y_{ab}$.

EXAMPLE 2 (Vector autoregressive processes). Recall that a vector autoregressive (VAR) process is defined by the recursion (2), and suppose that we observe an n-sequence $\{Z_t\}_{t=1}^n$ produced by this recursion. Since each $Z_t = [Z_{t1} \ \cdots \ Z_{tm}]^T$ is m-variate, the scalarized sample size is $N = nm$. Letting $b = 1, 2, \ldots, m$ index the dimension, we have

$$Z_{(t+1)b} = \langle\langle e_b Z_t^T, \Theta^* \rangle\rangle + W_{tb}. \tag{7}$$

In this case, we re-index the collection of $N = nm$ observations via the mapping

$$(t, b) \mapsto i = (t - 1)m + b.$$

After doing so, we see that the autoregressive problem can be written in the form (4) with $y_i = Z_{(t+1)b}$ and observation matrix $X_i = e_b Z_t^T$.

EXAMPLE 3 (Compressed sensing). As mentioned earlier, this is a natural extension of the compressed sensing observation model for sparse vectors to the case of low-rank matrices [45, 46]. In a typical form of compressed sensing, the observation matrix $X_i \in \mathbb{R}^{m_1 \times m_2}$ has i.i.d. standard normal $N(0, 1)$ entries, so that one makes observations of the form

$$y_i = \langle\langle X_i, \Theta^* \rangle\rangle + \varepsilon_i \qquad \text{for } i = 1, 2, \ldots, N. \tag{8}$$

By construction, these observations are an instance of the model (4). In the sequel, we study a more general observation model, in which the entries of X_i are allowed to have general Gaussian dependencies. For this problem, the more compact form (5) involves a random Gaussian operator mapping $\mathbb{R}^{m_1 \times m_2}$ to $\mathbb{R}^N$, and we study some of its properties in the sequel.

2.3. *Regression with nuclear norm regularization.* We now consider an estimator that is naturally suited to the problems described in the previous section. Recall that the *nuclear or trace norm* of a matrix $\Theta \in \mathbb{R}^{m_1 \times m_2}$ is given by $|||\Theta|||_1 = \sum_{j=1}^{m} \sigma_j(\Theta)$, corresponding to the sum of its singular values. Given a collection of observations $(y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{m_1 \times m_2}$, for $i = 1, \ldots, N$ from the observation model (4), we consider estimating the unknown $\Theta^* \in \mathcal{S}$ by solving the following optimization problem:

$$\widehat{\Theta} \in \arg\min_{\Theta \in \mathcal{S}} \left\{ \frac{1}{2N} \|\vec{y} - \mathfrak{X}(\Theta)\|_2^2 + \lambda_N |||\Theta|||_1 \right\}, \tag{9}$$

where $\mathcal{S}$ is a convex subset of $\mathbb{R}^{m_1 \times m_2}$, and $\lambda_N > 0$ is a regularization parameter. When $\mathcal{S} = \mathbb{R}^{m_1 \times m_2}$, the optimization problem (9) can be viewed as the analog of the Lasso estimator [50], tailored to low-rank matrices as opposed to sparse vectors. We include the possibility of a more general convex set $\mathcal{S}$ since they arise naturally in certain applications (e.g., Hankel-type constraints in system identification [33]). When $\mathcal{S}$ is a polytope (with $\mathcal{S} = \mathbb{R}^{m_1 \times m_2}$ as a special case), then the optimization problem (9) can be solved in time polynomial in the sample size N and the matrix dimensions m_1 and m_2. Indeed, the optimization problem (9) is an instance of a *semidefinite program* [51], a class of convex optimization problems that can be solved efficiently by various polynomial-time algorithms [11]. For instance, Liu and Vandenberghe [33] develop an efficient interior point method for solving constrained versions of nuclear norm programs. Moreover, as we discuss in Section 5, there are a variety of first-order methods for solving the semidefinite program (SDP) defining our M-estimator [29, 40]. These first-order methods are well suited to the high-dimensional problems arising in statistical settings, and we make use of one in performing our simulations.

Like in any typical M-estimator for statistical inference, the regularization parameter λ_N is specified by the statistician. As part of the theoretical results in the next section, we provide suitable choices of this parameter so that the estimate $\widehat{\Theta}$ is close in Frobenius norm to the unknown matrix Θ^*. The setting of the regularizer depends on the knowledge of the noise variance. While in general one might need to estimate this parameter through cross validation [9, 20], we assume knowledge of the noise variance in order to most succinctly demonstrate the empirical behavior of our results through the experiments.

3. Main results and some consequences. In this section, we state our main results and discuss some of their consequences. Section 3.1 is devoted to results that apply to generic instances of low-rank problems, whereas Section 3.3 is devoted to the consequences of these results for more specific problem classes, including low-rank multivariate regression, estimation of vector autoregressive processes and recovery of low-rank matrices from random projections.

3.1. *Results for general model classes.* We begin by introducing the key technical condition that allows us to control the error $\widehat{\Theta} - \Theta^*$ between an SDP solution $\widehat{\Theta}$ and the unknown matrix Θ^*. We refer to it as the *restricted strong convexity* condition [37], since it amounts to guaranteeing that the quadratic loss function in the SDP (9) is strictly convex over a restricted set of directions. Letting $\mathcal{C} \subseteq \mathbb{R}^{m_1 \times m_2}$ denote the restricted set of directions, we say that the operator $\mathfrak{X}$ satisfies restricted strong convexity (RSC) over the set $\mathcal{C}$ if there exists some $\kappa(\mathfrak{X}) > 0$ such that

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \geq \kappa(\mathfrak{X})|||\Delta|||_F^2 \qquad \text{for all } \Delta \in \mathcal{C}. \tag{10}$$

We note that analogous conditions have been used to establish error bounds in the context of sparse linear regression [10, 17], in which case the set $\mathcal{C}$ corresponded to certain subsets of sparse vectors. These types of conditions are weaker than restricted isometry properties, since they involve only lower bounds on the operator $\mathfrak{X}$, and the constant $\kappa(\mathfrak{X})$ can be arbitrarily small.

Of course, the definition (10) hinges on the choice of the restricted set $\mathcal{C}$. In order to specify some appropriate sets for the case of (near) low-rank matrices, we require some additional notation. Any matrix $\Theta^* \in \mathbb{R}^{m_1 \times m_2}$ has a singular value decomposition of the form $\Theta^* = UDV^T$, where $U \in \mathbb{R}^{m_1 \times m_1}$ and $V \in \mathbb{R}^{m_2 \times m_2}$ are orthonormal matrices. For each integer $r \in \{1, 2, \ldots, m\}$, we let $U^r \in \mathbb{R}^{m_1 \times r}$ and $V^r \in \mathbb{R}^{m_2 \times r}$ be the sub-matrices of singular vectors associated with the top r singular values of Θ^*. We then define the following two subspaces of $\mathbb{R}^{m_1 \times m_2}$:

$$\mathcal{A}(U^r, V^r) := \{\Delta \in \mathbb{R}^{m_1 \times m_2} \mid \text{row}(\Delta) \subseteq V^r \text{ and } \text{col}(\Delta) \subseteq U^r\} \tag{11a}$$

and

$$\mathcal{B}(U^r, V^r) := \{\Delta \in \mathbb{R}^{m_1 \times m_2} \mid \text{row}(\Delta) \perp V^r \text{ and } \text{col}(\Delta) \perp U^r\}, \tag{11b}$$

where $\text{row}(\Delta) \subseteq \mathbb{R}^{m_2}$ and $\text{col}(\Delta) \subseteq \mathbb{R}^{m_1}$ denote the row space and column space, respectively, of the matrix Δ. When (U^r, V^r) are clear from the context, we adopt the shorthand notation $\mathcal{A}^r$ and $\mathcal{B}^r$.

We can now define the subsets of interest. Let $\Pi_{\mathcal{B}^r}$ denote the projection operator onto the subspace $\mathcal{B}^r$, and define $\Delta'' = \Pi_{\mathcal{B}^r}(\Delta)$ and $\Delta' = \Delta - \Delta''$. For a positive integer $r \leq m = \min\{m_1, m_2\}$ and a tolerance parameter $\delta \geq 0$, consider the following subset of matrices:

$$\mathcal{C}(r; \delta) := \Bigg\{ \Delta \in \mathbb{R}^{m_1 \times m_2} \mid |||\Delta|||_F \geq \delta, \; |||\Delta''|||_1 \leq 3|||\Delta'|||_1 + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*) \Bigg\}. \tag{12}$$

Note that this set corresponds to matrices Δ for which the quantity $|||\Delta''|||_1$ is relatively small compared to $\Delta - \Delta''$ and the remaining $m - r$ singular values of Θ^*.

The next ingredient is the choice of the regularization parameter λ_N used in solving the SDP (9). Our theory specifies a choice for this quantity in terms of the adjoint of the operator $\mathfrak{X}$—namely, the operator $\mathfrak{X}^* : \mathbb{R}^N \to \mathbb{R}^{m_1 \times m_2}$ defined by

$$\mathfrak{X}^*(\vec{\varepsilon}) := \sum_{i=1}^{N} \varepsilon_i X_i. \tag{13}$$

With this notation, we come to the first result of our paper. It is a deterministic result, which specifies two conditions—namely, an RSC condition and a choice of the regularizer—that suffice to guarantee that any solution of the SDP (9) falls within a certain radius.

THEOREM 1. *Suppose $\Theta^* \in \mathcal{S}$ and that the operator $\mathfrak{X}$ satisfies restricted strong convexity with parameter $\kappa(\mathfrak{X}) > 0$ over the set $\mathcal{C}(r; \delta)$, and that the regularization parameter λ_N is chosen such that $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}/N$. Then any solution $\widehat{\Theta}$ to the semidefinite program (9) satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\left\{\delta, \frac{32\lambda_N\sqrt{r}}{\kappa(\mathfrak{X})}, \left[\frac{16\lambda_N \sum_{j=r+1}^{m} \sigma_j(\Theta^*)}{\kappa(\mathfrak{X})}\right]^{1/2}\right\}. \tag{14}$$

Apart from the tolerance parameter δ, the two main terms in the bound (14) have a natural interpretation. The first term (involving $\sqrt{r}$) corresponds to *estimation error*, capturing the difficulty of estimating a rank r matrix. The second is an *approximation error* that describes the gap between the true matrix Θ^* and the best rank r approximation. Understanding the magnitude of the tolerance parameter δ is a bit more subtle, and it depends on the geometry of the set $\mathcal{C}(r; \delta)$ and, more specifically, the inequality

$$|||\Delta''|||_1 \leq 3|||\Delta'|||_1 + 4 \sum_{j=r+1}^{m} \sigma_j(\Theta^*). \tag{15}$$

In the simplest case, when Θ^* is at most rank r, then we have $\sum_{j=r+1}^{m} \sigma_j(\Theta^*) = 0$, so the constraint (15) defines a cone. This cone completely excludes certain directions, and thus it is possible that the operator $\mathfrak{X}$, while failing RSC in a global sense, can satisfy it over the cone. Therefore, there is no need for a nonzero tolerance parameter δ in the exact low-rank case. In contrast, when Θ^* is only approximately low-rank, then the constraint (15) no longer defines a cone; rather, it includes an open ball around the origin. Thus, if $\mathfrak{X}$ fails RSC in a global sense, then it will also fail it under the constraint (15). The purpose of the additional constraint $|||\Delta|||_F \geq \delta$ is to eliminate the open ball centered at the origin, so that it is possible that $\mathfrak{X}$ satisfies RSC over $\mathcal{C}(r, \delta)$.

Let us now illustrate the consequences of Theorem 1 when the true matrix Θ^* has exactly rank r, in which case the approximation error term is zero. For the technical reasons mentioned above, it suffices to set $\delta = 0$ in the case of exact rank constraints, and we thus obtain the following result:

COROLLARY 1 (Exact low-rank recovery). *Suppose that $\Theta^* \in \mathcal{S}$ has rank r, and $\mathfrak{X}$ satisfies RSC with respect to $\mathcal{C}(r; 0)$. Then as long as $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}/N$, any optimal solution $\widehat{\Theta}$ to the SDP (9) satisfies the bound*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \frac{32\sqrt{r}\lambda_N}{\kappa(\mathfrak{X})}. \tag{16}$$

Like Theorem 1, Corollary 1 is a deterministic statement on the SDP error. It takes a much simpler form since when Θ^* is exactly low rank, then neither tolerance parameter δ nor the approximation term are required.

As a more delicate example, suppose instead that Θ^* is *nearly low-rank*, an assumption that we can formalize by requiring that its singular value sequence $\{\sigma_i(\Theta^*)\}_{i=1}^m$ decays quickly enough. In particular, for a parameter $q \in [0, 1]$ and a positive radius R_q, we define the set

$$\mathbb{B}_q(R_q) := \left\{ \Theta \in \mathbb{R}^{m_1 \times m_2} \;\Big|\; \sum_{i=1}^{m} |\sigma_i(\Theta)|^q \leq R_q \right\}, \tag{17}$$

where $m = \min\{m_1, m_2\}$. Note that when $q = 0$, the set $\mathbb{B}_0(R_0)$ corresponds to the set of matrices with rank at most R_0.

COROLLARY 2 (Near low-rank recovery). *Suppose that $\Theta^* \in \mathbb{B}_q(R_q) \cap \mathcal{S}$, the regularization parameter is lower bounded as $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}/N$, and the operator $\mathfrak{X}$ satisfies RSC with parameter $\kappa(\mathfrak{X}) \in (0, 1]$ over the set $\mathcal{C}(R_q\lambda_N^{-q}; \delta)$. Then any solution $\widehat{\Theta}$ to the SDP (9) satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\left\{\delta, 32\sqrt{R_q}\left(\frac{\lambda_N}{\kappa(\mathfrak{X})}\right)^{1-q/2}\right\}. \tag{18}$$

Note that the error bound (18) reduces to the exact low rank case (16) when $q = 0$ and $\delta = 0$. The quantity $\lambda_N^{-q}R_q$ acts as the "effective rank" in this setting, as clarified by our proof in Section 4.2. This particular choice is designed to provide an optimal trade-off between the approximation and estimation error terms in Theorem 1. Since λ_N is chosen to decay to zero as the sample size N increases, this effective rank will increase, reflecting the fact that as we obtain more samples, we can afford to estimate more of the smaller singular values of the matrix Θ^*.

3.2. *Comparison to related work.* Past work by Lee and Bresler [32] provides stability results on minimizing the nuclear norm with a quadratic constraint, or equivalently, performing least-squares with nuclear norm constraints. Their results are based on the restricted isometry property (RIP), which is more restrictive than than the RSC condition given here; see Examples 4 and 5 for concrete examples of operators $\mathfrak{X}$ that satisfy RSC but fail RIP. In our notation, their stability results guarantee that the error $|||\widehat{\Theta} - \Theta^*|||_F$ is bounded by a quantity proportional

$t := \|y - \mathfrak{X}(\Theta^*)\|_2/\sqrt{N}$. Given the observation model (5) with a noise vector $\vec{\varepsilon}$ in which each entry is i.i.d., zero mean with variance ν^2, note that we have $t \approx \nu$ with high probability. Thus, although such a result guarantees stability, it does not guarantee consistency, since for any fixed noise variance $\nu^2 > 0$, the error bound does not tend to zero as the sample size N increases. In contrast, our bounds all depend on the noise and sample size via the regularization parameter, whose optimal choice is $\lambda_N^* = 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}/N$. As will be clarified in Corollaries 3 through 5 to follow, for noise $\vec{\varepsilon}$ with variance ν and various choices of $\mathfrak{X}$, this regularization parameter satisfies the scaling $\lambda_N^* \asymp \nu\sqrt{\frac{m_1+m_2}{N}}$. Thus, our results guarantee consistency of the estimator, meaning that the error tends to zero as the sample size increases.

As previously noted, some concurrent work [13, 47] has also provided results on estimation of high-dimensional matrices in the noisy and statistical setting. Rohde and Tsybakov [47] derive results for estimating low-rank matrices based on a quadratic loss term regularized by the Schatten-q norm for $0 < q \leq 1$. Note that the nuclear norm ($q = 1$) is a convex program, whereas the values $q \in (0, 1)$ provide analogs on concave regularized least squares [20] in the linear regression setting. They provide results on both multivariate regression and matrix completion; most closely related to our work are the results on multivariate regression, which we discuss at more length following Corollary 3 below. On the other hand, Candes and Plan [13] present error rates in the Frobenius norm for estimating approximately low-rank matrices under the compressed sensing model, and we discuss below the connection to our Corollary 5 for this particular observation model. A major difference between our work and this body of work lies in the assumptions imposed on the observation operator $\mathfrak{X}$. All of the papers [13, 32, 47] impose the restricted isometry property (RIP), which requires that all restricted singular values of $\mathfrak{X}$ very close to 1 (so that it is a near-isometry). In contrast, we require only the restricted strong convexity (RSC) condition, which imposes only an arbitrarily small but positive lower bound on the operator. It is straightforward to construct operators $\mathfrak{X}$ that satisfy RSC while failing RIP, as we discuss in Examples 4 and 5 to follow.

3.3. *Results for specific model classes.* As stated, Corollaries 1 and 2 are fairly abstract in nature. More importantly, it is not immediately clear how the key underlying assumption—namely, the RSC condition—can be verified, since it is specified via subspaces that depend on Θ^*, which is, itself, the unknown quantity that we are trying to estimate. Nonetheless, we now show how, when specialized to more concrete models, these results yield concrete and readily interpretable results. As will be clear in the proofs of these results, each corollary requires overcoming two main technical obstacles: establishing that the appropriate form of the RSC property holds in a uniform sense (so that a priori knowledge of Θ^* is not required) and specifying an appropriate choice of the regularization parameter λ_N.

Each of these two steps is nontrivial, requiring some random matrix theory, but the end results are simply stated upper bounds that hold with high probability.

We begin with the case of rank-constrained multivariate regression. As discussed earlier in Example 1, recall that we observe pairs $(Y_i, Z_i) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ linked by the linear model $Y_i = \Theta^* Z_i + W_i$, where $W_i \sim N(0, \nu^2 I_{m_1 \times m_1})$ is observation noise. Here we treat the case of *random design regression*, meaning that the covariates Z_i are modeled as random. In particular, in the following result, we assume that $Z_i \sim N(0, \Sigma)$, i.i.d. for some m_2-dimensional covariance matrix $\Sigma \succ 0$. Recalling that $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ denote the maximum and minimum eigenvalues, respectively, we have the following.

COROLLARY 3 (Low-rank multivariate regression). *Consider the random design multivariate regression model where $\Theta^* \in \mathbb{B}_q(R_q) \cap \mathcal{S}$. There are universal constants $\{c_i, i = 1, 2, 3\}$ such that if we solve the SDP (9) with regularization parameter $\lambda_N = 10 \frac{\nu}{m_1} \sqrt{\sigma_{\max}(\Sigma)} \sqrt{\frac{(m_1+m_2)}{n}}$, we have*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1 \left(\frac{\nu^2 \sigma_{\max}(\Sigma)}{\sigma_{\min}^2(\Sigma)} \right)^{1-q/2} R_q \left(\frac{m_1 + m_2}{n} \right)^{1-q/2} \tag{19}$$

with probability greater than $1 - c_2 \exp(-c_3(m_1 + m_2))$.

REMARKS. Corollary 3 takes a particularly simple form when $\Sigma = I_{m_2 \times m_2}$: then there exists a constant c_1' such that $|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1' \nu^{2-q} R_q (\frac{m_1+m_2}{n})^{1-q/2}$. When Θ^* is exactly low rank—that is, $q = 0$, and Θ^* has rank $r = R_0$—this simplifies even further to

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1' \frac{\nu^2 r(m_1 + m_2)}{n}.$$

The scaling in this error bound is easily interpretable: naturally, the squared error is proportional to the noise variance ν^2, and the quantity $r(m_1 + m_2)$ counts the number of degrees of freedom of a $m_1 \times m_2$ matrix with rank r. Note that if we did not impose any constraints on Θ^*, then since a $m_1 \times m_2$ matrix has a total of $m_1 m_2$ free parameters, we would expect at best[3] to obtain rates of the order $|||\widehat{\Theta} - \Theta^*|||_F^2 = \Omega(\frac{\nu^2 m_1 m_2}{n})$. Note that when Θ^* is low rank—in particular, when $r \ll \min\{m_1, m_2\}$—then the nuclear norm estimator achieves substantially faster rates.[4]

[3]To clarify our use of sample size, we can either view the multivariate regression model as consisting of n samples with a constant SNR, or as N samples with SNR of order $1/m_1$. We adopt the former interpretation here.

[4]We also note that as stated, the result requires that $(m_1 + m_2)$ tend to infinity in order for the claim to hold with high probability. Although such high-dimensional scaling is the primary focus of this paper, we note that for application to the classical setting of fixed (m_1, m_2), the same statement (with different constants) holds with $m_1 + m_2$ replaced by $\log n$.

It is worth comparing this corollary to a result on multivariate regression due to Rohde and Tsybakov [47]. Their result applies to exactly low-rank matrices (say with rank r), but provides bounds on general Schatten norms (including the Frobenius norm). In this case, it provides a comparable rate when we make the setting $q = 0$ and $R_0 = r$ in the bound (19), namely showing that we require roughly $n \approx r(m_1 + m_2)$ samples, corresponding to the number of degrees of freedom. A significant difference lies in the conditions imposed on the design matrices: whereas their result is derived under RIP conditions on the design matrices, we require only the milder RSC condition. The following example illustrates the distinction for this model.

EXAMPLE 4 (Failure of RIP for multivariate regression). Under the random design model for multivariate regression, we have

$$F(\Theta) := \frac{\mathbb{E}[\|\mathfrak{X}(\Theta)\|_2^2]}{n|||\Theta|||_F^2} = \frac{\sum_{j=1}^{m_2} \|\sqrt{\Sigma}\Theta_j\|_2^2}{|||\Theta|||_F^2}, \tag{20}$$

where Θ_j is the jth row of Θ. In order for RIP to hold, it is necessary that quantity $F(\Theta)$ is extremely close to 1—certainly less than two—for all low-rank matrices. We now show that this cannot hold unless Σ has a small condition number. Let $v \in \mathbb{R}^{m_2}$ and $v' \in \mathbb{R}^{m_2}$ denote the minimum and maximum eigenvectors of Σ. By setting $\Theta = e_1 v^T$, we obtain a rank one matrix for which $F(\Theta) = \sigma_{\min}(\Sigma)$, and similarly, setting $\Theta' = e_1 (v')^T$ yields another rank one matrix for which $F(\Theta') = \sigma_{\max}(\Sigma)$. The preceding discussion applies to the average $\mathbb{E}[\|\mathfrak{X}(\Theta)\|_2^2]/n$, but since the individual matrices matrices X_i are i.i.d. and Gaussian, we have

$$\frac{\|\mathfrak{X}(\Theta)\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^{n} \langle\langle X_i, \Theta\rangle\rangle^2 \le 2F(\Theta) = 2\sigma_{\min}(\Sigma)$$

with high probability, using χ^2-tail bounds. Similarly, $\|\mathfrak{X}(\Theta')\|_2^2/n \ge (1/2) \times \sigma_{\max}(\Sigma)$ with high probability. Thus, we have exhibited a pair of rank one matrices with $|||\Theta|||_F = |||\Theta'|||_F = 1$ for which

$$\frac{\|\mathfrak{X}(\Theta')\|_2^2}{\|\mathfrak{X}(\Theta)\|_2^2} \ge \frac{1}{4}\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)}.$$

Consequently, unless $\sigma_{\max}(\Sigma)/\sigma_{\min}(\Sigma) \le 64$, it is not possible for RIP to hold with constant $\delta \le 1/2$. In contrast, as our results show, the RSC will hold w.h.p. whenever $\sigma_{\min}(\Sigma) > 0$, and the error is allowed to scale with the ratio $\sigma_{\max}(\Sigma)/\sigma_{\min}(\Sigma)$.

Next we turn to the case of estimating the system matrix Θ^* of an autoregressive (AR) model, as discussed in Example 2.

COROLLARY 4 (Autoregressive models). *Suppose that we are given n samples $\{Z_t\}_{t=1}^n$ from a m-dimensional autoregressive process (2) that is stationary, based on a system matrix that is stable ($|||\Theta^*|||_{\mathrm{op}} \leq \gamma < 1$) and approximately low-rank ($\Theta^* \in \mathbb{B}_q(R_q) \cap \mathcal{S}$). Then there are universal constants $\{c_i, i = 1, 2, 3\}$ such that if we solve the SDP (9) with regularization parameter $\lambda_N = \frac{2c_0|||\Sigma|||_{\mathrm{op}}}{m(1-\gamma)}\sqrt{\frac{m}{n}}$, then any solution $\widehat{\Theta}$ satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1 \left[\frac{\sigma_{\max}^2(\Sigma)}{\sigma_{\min}^2(\Sigma)(1-\gamma)^2}\right]^{1-q/2} R_q \left(\frac{m}{n}\right)^{1-q/2} \tag{21}$$

with probability greater than $1 - c_2 \exp(-c_3 m)$.

REMARKS. Like Corollary 3, the result as stated requires that the matrix dimension m tends to infinity, but the same bounds hold with m replaced by $\log n$, yielding results suitable for classical (fixed dimension) scaling. Second, the factor $(m/n)^{1-q/2}$, like the analogous term[5] in Corollary 3, shows that faster rates are obtained if Θ^* can be well approximated by a low rank matrix, namely for choices of the parameter $q \in [0, 1]$ that are closer to zero. Indeed, in the limit $q = 0$, we again reduce to the case of an exact rank constraint $r = R_0$, and the corresponding squared error scales as rm/n. In contrast to the case of multivariate regression, the error bound (21) also depends on the upper bound $|||\Theta^*|||_{\mathrm{op}} = \gamma < 1$ on the operator norm of the system matrix Θ^*. Such dependence is to be expected since the quantity γ controls the (in)stability and mixing rate of the autoregressive process. As clarified in the proof, the dependence of the sampling in the AR model also presents some technical challenges not present in the setting of multivariate regression.

Finally, we turn to the analysis of the compressed sensing model for matrix recovery, as initially described in Example 3. Although standard compressed sensing is based on observation matrices X_i with i.i.d. elements, here we consider a more general model that allows for dependence between the entries of X_i. First defining the quantity $M = m_1 m_2$, we use $\mathrm{vec}(X_i) \in \mathbb{R}^M$ to denote the vectorized form of the $m_1 \times m_2$ matrix X_i. Given a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{M \times M}$, we say that the observation matrix X_i is sampled from the Σ-ensemble if $\mathrm{vec}(X_i) \sim N(0, \Sigma)$. Finally, we define the quantity

$$\rho^2(\Sigma) := \sup_{\|u\|_2=1, \|v\|_2=1} \mathrm{var}(u^T X v), \tag{22}$$

where the random matrix $X \in \mathbb{R}^{m_1 \times m_2}$ is sampled from the Σ-ensemble. In the special case $\Sigma = I$, corresponding to the usual compressed sensing model, we have $\rho^2(I) = 1$.

[5]The term in Corollary 3 has a factor $m_1 + m_2$, since the matrix in that case could be nonsquare in general.

The following result applies to any observation model in which the noise vector $\vec{\varepsilon} \in \mathbb{R}^N$ satisfies the bound $\|\vec{\varepsilon}\|_2 \leq 2\nu\sqrt{N}$ for some constant ν. This assumption that holds for any bounded noise, and also holds with high probability for any random noise vector with sub-Gaussian entries with parameter ν. [The simplest example is that of Gaussian noise $N(0, \nu^2)$.]

COROLLARY 5 (Compressed sensing with dependent sampling). *Suppose that the matrices $\{X_i\}_{i=1}^N$ are drawn i.i.d. from the Σ-Gaussian ensemble, and that the unknown matrix $\Theta^* \in \mathbb{B}_q(R_q) \cap \mathcal{S}$ for some $q \in (0, 1]$. Then there are universal constants c_i such that for a sample size $N > c_1 \rho^2(\Sigma) R_q^{1-q/2}(m_1 + m_2)$, any solution $\widehat{\Theta}$ to the SDP (9) with regularization parameter $\lambda_N = c_0 \rho(\Sigma)\nu\sqrt{\frac{m_1+m_2}{N}}$ satisfies the bound*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_2 R_q \left(\frac{(\nu^2 \vee 1)(\rho^2(\Sigma)/\sigma_{\min}^2(\Sigma))(m_1 + m_2)}{N} \right)^{1-q/2} \tag{23}$$

with probability greater than $1 - c_3 \exp(-c_4(m_1 + m_2))$. In the special case $q = 0$ and Θ^* of rank r, we have

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_2 \frac{\rho^2(\Sigma)\nu^2}{\sigma_{\min}^2(\Sigma)} \frac{r(m_1 + m_2)}{N}. \tag{24}$$

The central challenge in proving this result is in proving an appropriate form of the RSC property. The following result on the random operator $\mathfrak{X}$ may be of independent interest here:

PROPOSITION 1. *Consider the random operator $\mathfrak{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^N$ formed by sampling from the Σ-ensemble. Then it satisfies*

$$\frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{N}} \geq \frac{1}{4} \|\sqrt{\Sigma}\,\mathrm{vec}(\Theta)\|_2 - 12\rho(\Sigma)\left(\sqrt{\frac{m_1}{N}} + \sqrt{\frac{m_2}{N}}\right)|||\Theta|||_1 \quad \text{for all } \Theta \in \mathbb{R}^{m_1 \times m_2} \tag{25}$$

with probability at least $1 - 2\exp(-N/32)$.

The proof of this result, provided in Appendix H, makes use of the Gordon–Slepian inequalities for Gaussian processes, and concentration of measure. As we show in Section C, it implies the form of the RSC property needed to establish Corollary 5.

In concurrent work, Candes and Plan [13] derived a result similar to Corollary 5 for the compressed sensing observation model. Their result applies to matrices with i.i.d. elements with sub-Gaussian tail behavior. While the analysis given here is specific to Gaussian random matrices, it allows for general dependence among

the entries. Their result applies only under certain restrictions on the sample size relative to matrix dimension and rank, whereas our result holds more generally without these extra conditions. Moreover, their proof relies on an application of RIP, which is in general more restrictive than the RSC condition used in our analysis. The following example provides a concrete illustration of a matrix family where the restricted isometry constants are unbounded as the rank r grows, but RSC still holds.

EXAMPLE 5 (RSC holds when RIP violated). Here we consider a family of random operators $\mathfrak{X}$ for which RSC holds with high probability, while RIP fails. Consider generating an i.i.d. collection of design matrices $X_i \in \mathbb{R}^{m\times m}$, each of the form

$$X_i = z_i I_{m\times m} + G_i \qquad \text{for } i = 1, 2, \ldots, N, \tag{26}$$

where $z_i \sim N(0, 1)$ and $G_i \in \mathbb{R}^{m\times m}$ is a standard Gaussian random matrix, independent of z_i. Note that we have $\operatorname{vec}(X_i) \sim N(0, \Sigma)$, where the $m^2 \times m^2$ covariance matrix has the form

$$\Sigma = \operatorname{vec}(I_{m\times m}) \operatorname{vec}(I_{m\times m})^T + I_{m^2\times m^2}. \tag{27}$$

Let us compute the quantity $\rho(\Sigma) = \sup_{\|u\|_2=1, \|v\|_2=1} \operatorname{var}(u^T X v)$. By the independence of z and G in the model (26), we have

$$\rho(\Sigma) \le \operatorname{var}(z) \sup_{u\in S^{m_1-1}, v\in S^{m_2-1}} u^T v + \sup_{u\in S^{m_1-1}, v\in S^{m_2-1}} \operatorname{var}(u^T G v) \le 2.$$

Letting $\mathfrak{X}$ be the associated random operator, we observe that for any $\Theta \in \mathbb{R}^{m\times m}$, the independence of z_i and G_i implies that

$$\mathbb{E}\left[\frac{\|\mathfrak{X}(\Theta)\|_2^2}{N}\right] = \left\|\sqrt{\Sigma}\operatorname{vec}(\Theta)\right\|_2^2 = \operatorname{trace}(\Theta)^2 + |||\Theta|||_F^2 \ge |||\Theta|||_F^2.$$

Consequently, Proposition 1 implies that

$$\frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{N}} \ge \frac{1}{4}|||\Theta|||_F - 48\sqrt{\frac{m}{N}}|||\Theta|||_1 \qquad \text{for all } \Theta \in \mathbb{R}^{m\times m} \tag{28}$$

with high probability. As mentioned previously, we show in Section C how this type of lower bound implies the RSC condition needed for our results.

On the other hand, the random operator can never satisfy RIP (with the rank r increasing), as the following calculation shows. In this context, RIP requires that bounds of the form

$$\frac{\|\mathfrak{X}(\Theta)\|_2^2}{N|||\Theta|||_F^2} \in [1-\delta, 1+\delta] \qquad \text{for all } \Theta \text{ with rank at most } r,$$

where $\delta \in (0, 1)$ is a constant independent of r. Note that the bound (28) implies that a *lower bound* of this form holds as long as $N = \Omega(rm)$. Moreover, this lower

bound cannot be substantially sharpened, since the trace term plays no role for matrices with zero diagonals.

We now show that no such upper bound can ever hold. For a rank $1 \leq r < m$, consider the $m \times m$ matrix of the form

$$\Gamma := \begin{bmatrix} I_{r\times r}/\sqrt{r} & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(m-r)} \end{bmatrix}.$$

By construction, we have $|||\Gamma|||_F = 1$ and $\operatorname{trace}(\Gamma) = \sqrt{r}$. Consequently, we have

$$\mathbb{E}\left[\frac{\|\mathfrak{X}(\Gamma)\|_2^2}{N}\right] = \operatorname{trace}(\Gamma)^2 + |||\Gamma|||_F^2 = r + 1.$$

The independence of the matrices X_i implies that $\frac{\|\mathfrak{X}(\Gamma)\|_2^2}{N}$ is sharply concentrated around this expected value, so that we conclude that

$$\frac{\|\mathfrak{X}(\Gamma)\|_2^2}{N|||\Gamma|||_F^2} \geq \frac{1}{2}[1 + r]$$

with high probability, showing that RIP cannot hold with upper and lower bounds of the same order.

Finally, we note that Proposition 1 also implies an interesting property of the null space of the operator $\mathfrak{X}$, one which can be used to establish a corollary about recovery of low-rank matrices when the observations are noiseless. In particular, suppose that we are given the noiseless observations $y_i = \langle\langle X_i, \Theta^* \rangle\rangle$ for $i = 1, \ldots, N$, and that we try to recover the unknown matrix Θ^* by solving the SDP

$$\min_{\Theta \in \mathbb{R}^{m_1 \times m_2}} |||\Theta|||_1 \qquad \text{such that } \langle\langle X_i, \Theta \rangle\rangle = y_i \qquad \text{for all } i = 1, \ldots, N. \tag{29}$$

This recovery procedure was studied by Recht and colleagues [45, 46] in the special case that X_i is formed of i.i.d. $N(0, 1)$ entries. Proposition 1 allows us to obtain a sharp result on recovery using this method for Gaussian matrices with general dependencies.

COROLLARY 6 (Exact recovery with dependent sampling). *Suppose that the matrices $\{X_i\}_{i=1}^N$ are drawn i.i.d. from the Σ-Gaussian ensemble, and that $\Theta^* \in \mathcal{S}$ has rank r. Given $N > c_0\rho^2(\Sigma)r(m_1 + m_2)$ noiseless samples, then with probability at least $1 - 2\exp(-N/32)$, the SDP (29) recovers the matrix Θ^* exactly.*

This result removes some extra logarithmic factors that were included in initial work [45] and provides the appropriate analog to compressed sensing results for sparse vectors [14, 18]. Note that (like in most of our results) we have made little effort to obtain good constants in this result: the important property is that the sample size N scales linearly in both r and $m_1 + m_2$. We refer the reader to Recht, Xu and Hassibi [46], who study the standard Gaussian model under the scaling $r = \Theta(m)$ and obtain sharp results on the constants.

4. Proofs. We now turn to the proofs of Theorem 1 and Corollary 1 through Corollary 4. Owing to space constraints, we leave the proofs of Corollaries 5 and 6 to the Appendix. However, we note that these proofs make use of Proposition 1 in order to establish the respective RSC or restricted null space conditions. In each case, we provide the primary steps in the main text, with more technical details stated as lemmas and proved in the Appendix.

4.1. *Proof of Theorem 1.* By the optimality of $\widehat{\Theta}$ and feasibility of Θ^* for the SDP (9), we have

$$\frac{1}{2N}\|\vec{y} - \mathfrak{X}(\widehat{\Theta})\|_2^2 + \lambda_N |||\widehat{\Theta}|||_1 \leq \frac{1}{2N}\|\vec{y} - \mathfrak{X}(\Theta^*)\|_2^2 + \lambda_N |||\Theta^*|||_1.$$

Defining the error matrix $\Delta = \Theta^* - \widehat{\Theta}$ and performing some algebra yields the inequality

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \leq \frac{1}{N}\langle \vec{\varepsilon}, \mathfrak{X}(\Delta)\rangle + \lambda_N\{|||\widehat{\Theta} + \Delta|||_1 - |||\widehat{\Theta}|||_1\}. \tag{30}$$

By definition of the adjoint and Hölder's inequality, we have

$$\frac{1}{N}|\langle \vec{\varepsilon}, \mathfrak{X}(\Delta)\rangle| = \frac{1}{N}|\langle \mathfrak{X}^*(\vec{\varepsilon}), \Delta\rangle| \leq \frac{1}{N}|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}|||\Delta|||_1. \tag{31}$$

By the triangle inequality, we have $|||\widehat{\Theta} + \Delta|||_1 - |||\widehat{\Theta}|||_1 \leq |||\Delta|||_1$. Substituting this inequality and the bound (31) into the inequality (30) yields

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \leq \frac{1}{N}|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}|||\Delta|||_1 + \lambda_N|||\Delta|||_1 \leq 2\lambda_N|||\Delta|||_1,$$

where the second inequality makes use of our choice $\lambda_N \geq \frac{2}{N}|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}$.

It remains to lower bound the term on the left-hand side, while upper bounding the quantity $|||\Delta|||_1$ on the right-hand side. The following technical result allows us to do so. Recall our earlier definition (11) of the sets $\mathcal{A}$ and $\mathcal{B}$ associated with a given subspace pair.

LEMMA 1. *Let $U^r \in \mathbb{R}^{m_1 \times r}$ and $V^r \in \mathbb{R}^{m_2 \times r}$ be matrices consisting of the top r left and right (respectively) singular vectors of Θ^*. Then there exists a matrix decomposition $\Delta = \Delta' + \Delta''$ of the error Δ such that*:

(a) *the matrix Δ' satisfies the constraint* $\mathrm{rank}(\Delta') \leq 2r$;
(b) *if $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}/N$, then the nuclear norm of Δ'' is bounded as*

$$|||\Delta''|||_1 \leq 3|||\Delta'|||_1 + 4\sum_{j=r+1}^{m}\sigma_j(\Theta^*). \tag{32}$$

See Appendix B for the proof of this claim. Using Lemma 1, we can complete the proof of the theorem. In particular, from the bound (32) and the RSC assumption, we find that for $|||\Delta|||_F \geq \delta$, we have

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \geq \kappa(\mathfrak{X})|||\Delta|||_F^2.$$

Using the triangle inequality together with inequality (32), we obtain

$$|||\Delta|||_1 \leq |||\Delta'|||_1 + |||\Delta''|||_1 \leq 4|||\Delta'|||_1 + 4\sum_{j=r+1}^{m}\sigma_j(\Theta^*).$$

From the rank constraint in Lemma 1(a), we have $|||\Delta'|||_1 \leq \sqrt{2r}|||\Delta'|||_F$. Putting together the pieces, we find that either $|||\Delta|||_F \leq \delta$ or

$$\kappa(\mathfrak{X})|||\Delta|||_F^2 \leq \max\left\{32\lambda_N\sqrt{r}|||\Delta|||_F, 16\lambda_N\sum_{j=r+1}^{m}\sigma_j(\Theta^*)\right\},$$

which implies that

$$|||\Delta|||_F \leq \max\left\{\delta, \frac{32\lambda_N\sqrt{r}}{\kappa(\mathfrak{X})}, \left(\frac{16\lambda_N\sum_{j=r+1}^{m}\sigma_j(\Theta^*)}{\kappa(\mathfrak{X})}\right)^{1/2}\right\}$$

as claimed.

4.2. *Proof of Corollary 2.* Let $m = \min\{m_1, m_2\}$. In this case, we consider the singular value decomposition $\Theta^* = UDV^T$, where $U \in \mathbb{R}^{m_1\times m}$ and $V \in \mathbb{R}^{m_2\times m}$ are orthogonal, and we assume that D is diagonal with the singular values in nonincreasing order $\sigma_1(\Theta^*) \geq \sigma_2(\Theta^*) \geq \cdots \sigma_m(\Theta^*) \geq 0$. For a parameter $\tau > 0$ to be chosen, we define

$$K := \{i \in \{1, 2, \ldots, m\} \mid \sigma_i(\Theta^*) > \tau\},$$

and we let U^K (resp., V^K) denote the $m_1 \times |K|$ (resp., the $m_2 \times |K|$) orthogonal matrix consisting of the first $|K|$ columns of U (resp., V). With this choice, the matrix $\Theta^*_{K^c} := \Pi_{\mathcal{B}^{|K|}}(\Theta^*)$ has rank at most $m - |K|$, with singular values $\{\sigma_i(\Theta^*), i \in K^c\}$. Moreover, since $\sigma_i(\Theta^*) \leq \tau$ for all $i \in K^c$, we have

$$|||\Theta^*_{K^c}|||_1 = \tau\sum_{i=|K|+1}^{m}\frac{\sigma_i(\Theta^*)}{\tau} \leq \tau\sum_{i=|K|+1}^{m}\left(\frac{\sigma_i(\Theta^*)}{\tau}\right)^q \leq \tau^{1-q}R_q.$$

On the other hand, we also have $R_q \geq \sum_{i=1}^{m}|\sigma_i(\Theta^*)|^q \geq |K|\tau^q$, which implies that $|K| \leq \tau^{-q}R_q$. From the general error bound with $r = |K|$, we obtain

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\left\{\delta, \frac{32\lambda_N\sqrt{R_q}\tau^{-q/2}}{\kappa(\mathfrak{X})}, \left[\frac{16\lambda_N\tau^{1-q}R_q}{\kappa(\mathfrak{X})}\right]^{1/2}\right\}.$$

Setting $\tau = \lambda_N/\kappa(\mathfrak{X})$ yields that

$$\begin{aligned}|||\widehat{\Theta} - \Theta^*|||_F &\leq \max\left\{\delta, \frac{32\lambda_N^{1-q/2}\sqrt{R_q}}{\kappa^{1-q/2}}, \left[\frac{16\lambda_N^{2-q}R_q}{\kappa^{2-q}}\right]^{1/2}\right\} \\ &= \max\left\{\delta, 32\sqrt{R_q}\left(\frac{\lambda_N}{\kappa(\mathfrak{X})}\right)^{1-q/2}\right\}\end{aligned}$$

as claimed.

4.3. *Proof of Corollary* 3. For the proof of this corollary, we adopt the following notation. We first define the three matrices

$$\begin{aligned}X &= \begin{bmatrix} Z_1^T \\ Z_2^T \\ \cdots \\ Z_n^T \end{bmatrix} \in \mathbb{R}^{n\times m_2}, \qquad Y = \begin{bmatrix} Y_1^T \\ Y_2^T \\ \cdots \\ Y_n^T \end{bmatrix} \in \mathbb{R}^{n\times m_1} \quad \text{and} \\ W &= \begin{bmatrix} W_1^T \\ W_2^T \\ \cdots \\ W_n^T \end{bmatrix} \in \mathbb{R}^{n\times m_1}.\end{aligned} \tag{33}$$

With this notation and using the relation $N = nm_1$, the SDP objective function (9) can be written as $\frac{1}{m_1}\{\frac{1}{2n}|||Y - X\Theta^T|||_F^2 + \lambda_n|||\Theta|||_1\}$, where we have defined $\lambda_n = \lambda_N m_1$.

In order to establish the RSC property for this model, some algebra shows that we need to establish a lower bound on the quantity

$$\frac{1}{2n}|||X\Delta|||_F^2 = \frac{1}{2n}\sum_{j=1}^{m}\|(X\Delta)_j\|_2^2 \geq \frac{\sigma_{\min}(X^TX)}{2n}|||\Delta|||_F^2,$$

where $\sigma_{\min}$ denotes the minimum eigenvalue. The following lemma follows by adapting known concentration results for random matrices (see [52] for details):

LEMMA 2. *Let $X \in \mathbb{R}^{n\times m}$ be a random matrix with i.i.d. rows sampled from a m-variate $N(0, \Sigma)$ distribution. Then for $n \geq 2m$, we have*

$$\mathbb{P}\left[\sigma_{\min}\left(\frac{1}{n}X^TX\right) \geq \frac{\sigma_{\min}(\Sigma)}{9}, \sigma_{\max}\left(\frac{1}{n}X^TX\right) \leq 9\sigma_{\max}(\Sigma)\right] \geq 1 - 4\exp(-n/2).$$

As a consequence, we have $\frac{\sigma_{\min}(X^TX)}{2n} \geq \frac{\sigma_{\min}(\Sigma)}{18}$ with probability at least $1 - 4\exp(-n)$ for all $n \geq 2m$, which establishes that the RSC property holds with $\kappa(\mathfrak{X}) = \frac{1}{20m_1}\sigma_{\min}(\Sigma)$.

Next we need to upper bound the quantity $|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}}$ for this model, so as to verify that the stated choice for λ_N is valid. Following some algebra, we find that

$$\frac{1}{n}|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\mathrm{op}} = \frac{1}{n}|||X^T W|||_{\mathrm{op}}.$$

The following lemma is proved in Appendix F:

LEMMA 3. *There are constants $c_i > 0$ such that*

$$\mathbb{P}\left[\left|\frac{1}{n}|||X^T W|||_{\mathrm{op}}\right| \geq 5\nu\sqrt{\sigma_{\max}(\Sigma)}\sqrt{\frac{m_1+m_2}{n}}\right] \leq c_1 \exp\bigl(-c_2(m_1+m_2)\bigr). \tag{34}$$

Using these two lemmas, we can complete the proof of Corollary 3. First, recalling the scaling $N = m_1 n$, we see that Lemma 3 implies that the choice $\lambda_n = 10\nu\sqrt{\sigma_{\max}(\Sigma)}\sqrt{\frac{m_1+m_2}{n}}$ satisfies the conditions of Corollary 2 with high probability. Lemma 2 shows that the RSC property holds with $\kappa(\mathfrak{X}) = \sigma_{\min}(\Sigma)/(20m_1)$, again with high probability. Consequently, Corollary 2 implies that

$$\begin{aligned}|||\widehat{\Theta} - \Theta^*|||_F^2 &\leq 32^2 R_q\left(10\nu\sqrt{\sigma_{\max}(\Sigma)}\sqrt{\frac{m_1+m_2}{n}}\frac{20}{\sigma_{\min}(\Sigma)}\right)^{2-q} \\ &= c_1\left(\frac{\nu^2\sigma_{\max}(\Sigma)}{\sigma_{\min}^2(\Sigma)}\right)^{1-q/2} R_q\left(\frac{m_1+m_2}{n}\right)^{1-q/2}\end{aligned}$$

with probability greater than $1 - c_2\exp(-c_3(m_1+m_2))$, as claimed.

4.4. *Proof of Corollary 4.* For the proof of this corollary, we adopt the notation

$$X = \begin{bmatrix} Z_1^T \\ Z_2^T \\ \cdots \\ Z_n^T \end{bmatrix} \in \mathbb{R}^{n\times m} \quad \text{and} \quad Y = \begin{bmatrix} Z_2^T \\ Z_2^T \\ \cdots \\ Z_{n+1}^T \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

Finally, we let $W \in \mathbb{R}^{n\times m}$ be a matrix where each row is sampled i.i.d. from the $N(0, C)$ distribution corresponding to the innovations noise driving the VAR process. With this notation, and using the relation $N = nm$, the SDP objective function (9) can be written as $\frac{1}{m}\{\frac{1}{2n}|||Y - X\Theta^T|||_F^2 + \lambda_n|||\Theta|||_1\}$, where we have defined $\lambda_n = \lambda_N m$. At a high level, the proof of this corollary is similar to that of Corollary 3, in that we use random matrix theory to establish the required RSC property and to justify the choice of λ_n, or equivalently λ_N. However, it is considerably more challenging, due to the dependence in the rows of the random matrices, and the cross-dependence between the two matrices X and W (which were independent in the setting of multivariate regression).

The following lemma provides the lower bound needed to establish RSC for the autoregressive model:

LEMMA 4. *The eigenspectrum of the matrix $X^T X/n$ is well controlled in terms of the stationary covariance matrix*: *in particular, as long as $n > c_3 m$, we have*

$$\sigma_{\max}\left(\left(\frac{1}{n}X^T X\right)\right) \overset{(a)}{\leq} \frac{24\sigma_{\max}(\Sigma)}{1-\gamma} \quad \text{and} \quad \sigma_{\min}\left(\left(\frac{1}{n}X^T X\right)\right) \overset{(b)}{\geq} \frac{\sigma_{\min}(\Sigma)}{4}, \tag{35}$$

both with probability greater than $1 - 2c_1 \exp(-c_2 m)$.

Thus, from the bound (35)(b), we see with the high probability, the RSC property holds with $\kappa(\mathfrak{X}) = \sigma_{\min}(\Sigma)/(4m_2)$ as long as $n > c_3 m$.

As before, in order to verify the choice of λ_n, we need to control the quantity $\frac{1}{n}|||X^T W|||_{\text{op}}$. The following inequality, proved in Appendix G.2, yields a suitable upper bound:

LEMMA 5. *There exist constants $c_i > 0$, independent of n, m, Σ etc. such that*

$$\mathbb{P}\left[\frac{1}{n}|||X^T W|||_{\text{op}} \geq \frac{c_0 |||\Sigma|||_{\text{op}}}{1-\gamma}\sqrt{\frac{m}{n}}\right] \leq c_2 \exp(-c_3 m). \tag{36}$$

From Lemma 5, we see that it suffices to choose $\lambda_n = \frac{2c_0|||\Sigma|||_{\text{op}}}{1-\gamma}\sqrt{\frac{m}{n}}$. With this choice, Corollary 2 of Theorem 1 yields that

$$|||\Theta - \Theta^*|||_F^2 \leq c_1 R_q \left[\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)(1-\gamma)}\right]^{2-q} \left(\frac{m}{n}\right)^{1-q/2}$$

with probability greater than $1 - c_2 \exp(-c_3 m)$, as claimed.

5. Experimental results. In this section, we report the results of various simulations that demonstrate the close agreement between the scaling predicted by our theory, and the actual behavior of the SDP-based M-estimator (9) in practice. In all cases, we solved the convex program (9) by using our own implementation in MATLAB of an accelerated gradient descent method which adapts a non-smooth convex optimization procedure [40] to the nuclear-norm [29]. We chose the regularization parameter λ_N in the manner suggested by our theoretical results; in doing so, we assumed knowledge of the noise variance ν^2. In practice, one would have to estimate such quantities from the data using methods such as cross-validation, as has been studied in the context of the Lasso, and we leave this as an interesting direction for future research.

We report simulation results for three of the running examples discussed in this paper: low-rank multivariate regression, estimation in vector autoregressive processes and matrix recovery from random projections (compressed sensing). In each case, we solved instances of the SDP for a square matrix $\Theta^* \in \mathbb{R}^{m \times m}$, where $m \in \{40, 80, 160\}$ for the first two examples, and $m \in \{20, 40, 80\}$ for the compressed sensing example. In all cases, we considered the case of exact low rank

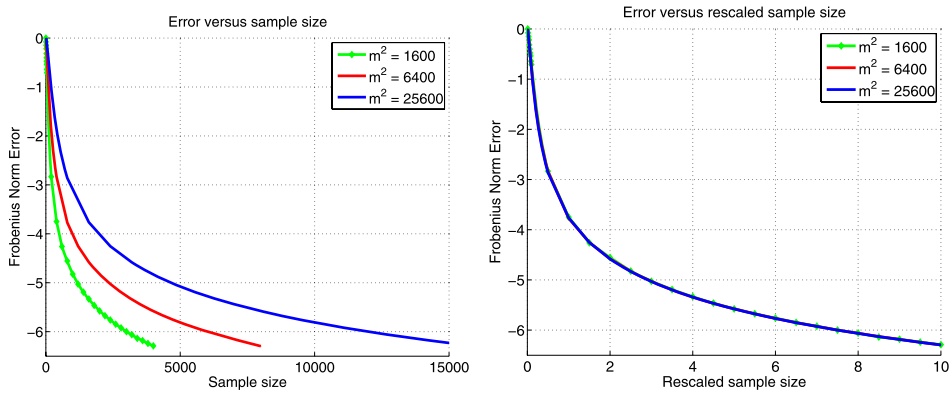


FIG. 1. *Results of applying the SDP* (9) *with nuclear norm regularization to the problem of low-rank multivariate regression.* (a) *Plots of the Frobenius error* $|||\widehat{\Theta} - \Theta^*|||_F$ *on a logarithmic scale versus the sample size* N *for three different matrix sizes* $m^2 \in \{1600, 6400, 25600\}$, *all with rank* $r = 10$. (b) *Plots of the same Frobenius error versus the rescaled sample size* $N/(rm)$. *Consistent with theory, all three plots are now extremely well aligned.*

constraints, with $\operatorname{rank}(\Theta^*) = r = 10$, and we generated Θ^* by choosing the subspaces of its left and right singular vectors uniformly at random from the Grassman manifold.[6] The observation noise had variance $\nu^2 = 1$, and we chose $C = \nu^2 I$ for the VAR process. The VAR process was generated by first solving for the covariance matrix Σ using the MATLAB function dylap and then generating a sample path. For each setting of (r, m), we solved the SDP for a range of sample sizes N.

Figure 1 shows results for a multivariate regression model with the covariates chosen randomly from a $N(0, I)$ distribution. Panel (a) plots the Frobenius error $|||\widehat{\Theta} - \Theta^*|||_F$ on a logarithmic scale versus the sample size N for three different matrix sizes, $m \in \{40, 80, 160\}$. Naturally, in each case, the error decays to zero as N increases, but larger matrices require larger sample sizes, as reflected by the rightward shift of the curves as m is increased. Panel (b) of Figure 1 shows the exact same set of simulation results, but now with the Frobenius error plotted versus the rescaled sample size $\widetilde{N} := N/(rm)$. As predicted by Corollary 3, the error plots now are all aligned with one another; the degree of alignment in this particular case is so close that the three plots are now indistinguishable. (The blue curve is the only one visible since it was plotted last by our routine.) Consequently, Figure 1 shows that $N/(rm)$ acts as the effective sample size in this high-dimensional setting.

Figure 2 shows similar results for the autoregressive model discussed in Example 2. As shown in panel (a), the Frobenius error again decays as the sample

[6]More specifically, we let $\Theta^* = XY^T$, where $X, Y \in \mathbb{R}^{m \times r}$ have i.i.d. $N(0, 1)$ elements.

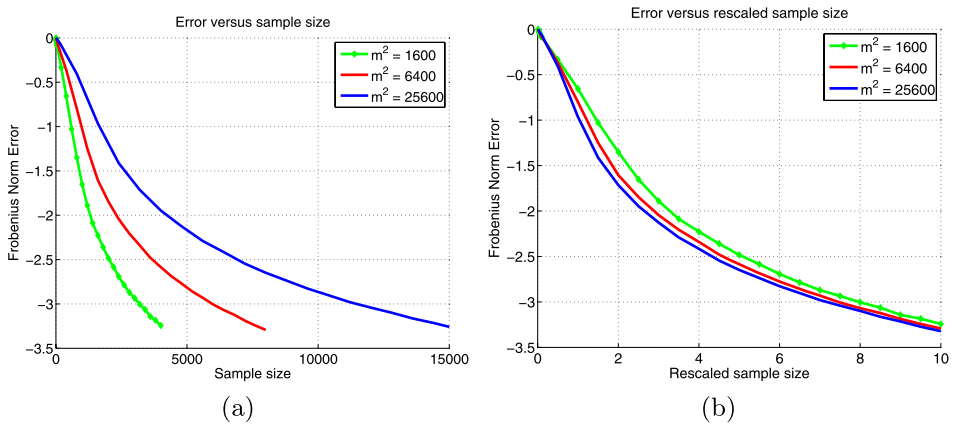
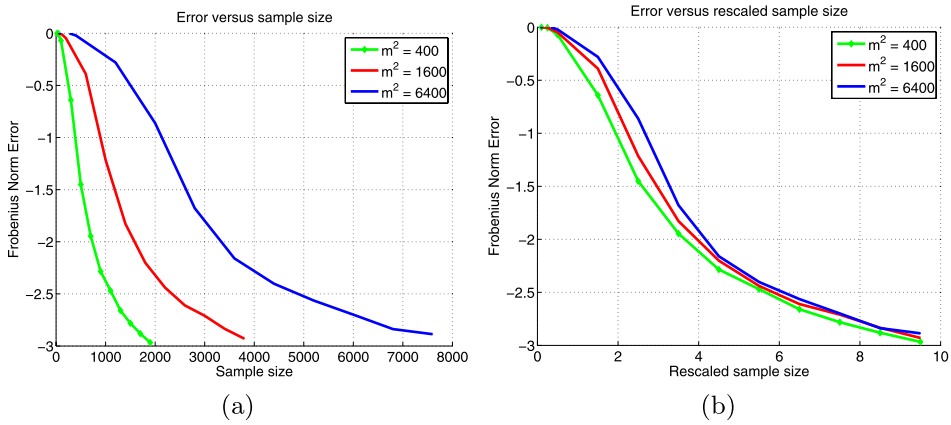


FIG. 2. *Results of applying the SDP* (9) *with nuclear norm regularization to estimating the system matrix of a vector autoregressive process*. (a) *Plots of the Frobenius error* $|||\widehat{\Theta} - \Theta^*|||_F$ *on a logarithmic scale versus the sample size* N *for three different matrix sizes* $m^2 \in \{1600, 6400, 25600\}$, *all with rank* $r = 10$. (b) *Plots of the same Frobenius error versus the rescaled sample size* $N/(rm)$. *Consistent with theory, all three plots are now reasonably well aligned.*

size is increased, although problems involving larger matrices are shifted to the right. Panel (b) shows the same Frobenius error plotted versus the rescaled sample size $N/(rm)$; as predicted by Corollary 4, the errors for different matrix sizes m are again quite well-aligned. In this case, we find (both in our theoretical analysis and experimental results) that the dependence in the autoregressive process slows down the rate at which the concentration occurs, so that the results are not as crisp as the low-rank multivariate setting in Figure 1.

Finally, Figure 3 presents the same set of results for the compressed sensing observation model discussed in Example 3. Even though the observation matrices X_i here are qualitatively different (in comparison to the multivariate regression and autoregressive examples), we again see the "stacking" phenomenon of the curves when plotted versus the rescaled sample size N/rm, as predicted by Corollary 5.

6. Discussion. In this paper, we have analyzed the nuclear norm relaxation for a general class of noisy observation models and obtained nonasymptotic error bounds on the Frobenius norm that hold under high-dimensional scaling. In contrast to most past work, our results are applicable to both exactly and approximately low-rank matrices. We stated a main theorem that provides high-dimensional rates in a fairly general setting, and then showed how by specializing this result to some specific model classes—namely, low-rank multivariate regression, estimation of autoregressive processes and matrix recovery from random projections—it yields concrete and readily interpretable rates. Finally, we provided some simulation results that showed excellent agreement with the predictions from



(a) (b)

FIG. 3. *Results of applying the SDP* (9) *with nuclear norm regularization to recovering a low-rank matrix on the basis of random projections* (*compressed sensing model*). (a) *Plots of the Frobenius error* $|||\widehat{\Theta} - \Theta^*|||_F$ *on a logarithmic scale versus the sample size* N *for three different matrix sizes* $m^2 \in \{400, 1600, 6400\}$, *all with rank* $r = 10$. (b) *Plots of the same Frobenius error versus the rescaled sample size* $N/(rm)$. *Consistent with theory, all three plots are now reasonably well aligned.*

our theory. Our more recent work has also shown that this same framework can be used to obtain near-optimal bounds for the matrix completion problem [38].

This paper has focused on achievable results for low-rank matrix estimation using a particular polynomial-time method. It would be interesting to establish matching lower bounds, showing that the rates obtained by this estimator are minimax-optimal. We suspect that this should be possible, for instance, by using the techniques exploited in Raskutti, Wainwright and Yu [43] in analyzing minimax rates for regression over ℓ_q-balls.

Acknowledgments. We thank the editors and anonymous reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation of (Near) Low-Rank Matrices with Noise and High-Dimensional Scaling" (DOI: 10.1214/10-AOS850SUPP; .pdf). Owing to space constraints, we have moved many of the technical proofs and details to the Appendix, which is contained in the supplementary document [39].

REFERENCES

[1] ABERNETHY, J., BACH, F., EVGENIOU, T. and STEIN, J. (2006). Low-rank matrix factorization with attributes. Technical Report N-24/06/MM, Ecole des mines de Paris, France.

[2] AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

[3] ANDERSON, C. W., STOLZ, E. A. and SHAMSUNDER, S. (1998). Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. *IEEE Trans. Bio-Med. Eng.* **45** 277.
[4] ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York. MR0283939
[5] ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2006). Multi-task feature learning. In *Neural Information Processing Systems* (*NIPS*) 41–48. Vancouver, Canada.
[6] BACH, F. (2008). Consistency of trace norm minimization. *J. Mach. Learn. Res.* **9** 1019–1048. MR2417263
[7] BICKEL, P. and LEVINA, E. (2008). Covariance estimation by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008
[8] BICKEL, P. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969
[9] BICKEL, P. and LI, B. (2006). Regularization in statistics. *TEST* **15** 271–344. MR2273731
[10] BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469
[11] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575
[12] BROWN, E. N., KASS, R. E. and MITRA, P. P. (2004). Multiple neural spike train data analysis: State-of-the-art and future challenges. *Nature Neuroscience* **7** 456–466.
[13] CANDÈS, E. and PLAN, Y. (2010). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Technical report, Stanford Univ. Available at arXiv:1001.0339v1.
[14] CANDES, E. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215. MR2243152
[15] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240
[16] CHEN, S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094
[17] COHEN, A., DAHMEN, W. and DEVORE, R. (2009). Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.* **22** 211–231. MR2449058
[18] DONOHO, D. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189
[19] EL-KAROUI, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011
[20] FAN, J. and LI, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581
[21] FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659
[22] FAZEL, M. (2002). *Matrix Rank Minimization with Applications*. Ph.D. thesis, Stanford Univ. Available at http://faculty.washington.edu/mfazel/thesis-final.pdf.
[23] FISHER, J. and BLACK, M. J. (2005). Motor cortical decoding using an autoregressive moving average model. *27th Annual International Conference of the Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005* 2130–2133.
[24] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.
[25] HARRISON, L., PENNY, W. D. and FRISTON, K. (2003). Multivariate autoregressive modeling of fmri time series. *NeuroImage* **19** 1477–1491.
[26] HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR0832183

[27] HORN, R. A. and JOHNSON, C. R. (1991). *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR1091716

[28] HUANG, J. and ZHANG, T. (2009). The benefit of group sparsity. Technical report, Rutgers Univ. Available at arXiv:0901.2962.

[29] JI, S. and YE, J. (2009). An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning* (*ICML*) 457–464. ACM, New York.

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

[31] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2009). Matrix completion from noisy entries. Technical report, Stanford Univ. Available at http://arxiv.org/abs/0906.2027v1.

[32] LEE, K. and BRESLER, Y. (2009). Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report. UIUC. Available at arXiv:0903.4742.

[33] LIU, Z. and VANDENBERGHE, L. (2009). Interior-point method for nuclear norm optimization with application to system identification. *SIAM J. Matrix Anal. Appl.* **31** 1235–1256. MR2558821

[34] LOUNICI, K., PONTIL, M., TSYBAKOV, A. B. and VAN DE GEER, S. (2009). Taking advantage of sparsity in multi-task learning. Technical report, ETH Zurich. Available at arXiv:0903.1468.

[35] LÜTKEPOLHL, H. (2006). *New Introduction to Multiple Time Series Analysis*. Springer, New York.

[36] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[37] 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 *Proceedings of the NIPS Conference* 1348–1356. Vancouver, Canada.

[38] NEGAHBAN, S. and WAINWRIGHT, M. J. (2010). Restricted strong convexity and (weighted) matrix completion: Near-optimal bounds with noise. Technical report, Univ. California, Berkeley.

[39] NEGAHBAN, S. and WAINWRIGHT, M. J. (2010). Supplement to "Estimation of (near) low-rank matrices with noise and high-dimensional scaling." DOI:10.1214/10-AOS850SUPP.

[40] NESTEROV, Y. (2007). Gradient methods for minimizing composite objective function. Technical Report 2007/76, CORE, Univ. Catholique de Louvain.

[41] OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Union support recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47.

[42] PAUL, D. and JOHNSTONE, I. (2008). Augmented sparse principal component analysis for high-dimensional data. Technical report, Univ. California, Davis.

[43] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2009). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical report, Dept. Statistics, Univ. California, Berkeley. Available at arXiv:0910.2042.

[44] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2008). High-dimensional covariance estimation: Convergence rates of ℓ_1-regularized log-determinant divergence. Technical report, Dept. Statistics, Univ. California, Berkeley.

[45] 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.

[46] RECHT, B., XU, W. and HASSIBI, B. (2009). Null space conditions and thresholds for rank minimization. Technical report, Univ. Wisconsin–Madison. Available at http://pages.cs.wisc.edu/~brecht/papers/10.RecXuHas.Thresholds.pdf.

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

[48] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic J. Statist.* **2** 494–515. MR2417391

[49] SREBRO, N., RENNIE, J. and JAAKKOLA, T. (2005). Maximum-margin matrix factorization. In *Proceedings of the NIPS Conference* 1329–1336. Vancouver, Canada.

[50] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[51] VANDENBERGHE, L. and BOYD, S. (1996). Semidefinite programming. *SIAM Rev.* **38** 49–95. MR1379041

[52] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202.

[53] 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

[54] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[55] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: sahand_n@eecs.berkeley.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@eecs.berkeley.edu