Estimation of (near) low-rank matrices with noise and high-dimensional scaling

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

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

Abstract

We study an instance of high-dimensional statistical inference in which the goal is to use N noisy observations to estimate a matrix $\Theta^* \in \mathbb{R}^{k \times p}$ that 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 an M-estimator based on regularization by the trace or nuclear norm over matrices, and analyze its performance under high-dimensional scaling. We provide non-asymptotic 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 their consequences for a number of specific learning models, including low-rank multivariate or multi-task regression, system identification in vector autoregressive processes, and recovery of low-rank matrices from random projections. Simulations 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 p" 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 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 (e.g., (1; 2; 3)). A theme common to much of this work is the use of ℓ_1-penalty as a surrogate function to enforce the sparsity constraint.

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 $\Theta^* \in \mathbb{R}^{k \times p}$ that is either *exactly low rank*, meaning that it has at most $r \ll \min\{k, p\}$ non-zero 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 ap-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

proximation has been studied by various researchers in optimization and machine learning (e.g., (4; 5; 6; 7; 8)), with the nuclear norm relaxation studied for (among other problems) noiseless random projections (9), as well as for matrix completion problems (10; 11). In addition, Bach (7) has provided some consistency results for nuclear norm regularization in the classical (fixed "p") setting, but not in the high-dimensional setting considered here.

The goal of this paper is to analyze the nuclear norm relaxation for a general class of noisy observation models, and obtain non-asymptotic 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 several cases of interest, including low-rank multivariate regression, estimation of autoregressive processes, and random projection (compressed sensing) observations. Our theoretical results on these models are obtained by leveraging the ideas from our own past work (12) on M-estimators with decomposable regularizers, where it is shown that error rates can be obtained by bounding the restricted strong convexity (RSC) parameter and specifying a suitable choice of the regularization parameter. Establishing bounds on these parameters for specific models can involve some non-trivial analysis, and in this paper, we use random matrix theory to provide the requisite control.

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 $\langle\langle \Theta, \Gamma \rangle\rangle = \mathrm{trace}(\Theta^T \Gamma)$ denote the trace inner product on matrix space. For a matrix $\Theta \in \mathbb{R}^{k \times p}$, we let $m = \min\{k, p\}$, and denote its (ordered) singular values by $\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \ldots \geq \sigma_m(\Theta) \geq 0$. 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* $|||\Theta|||_F = \sqrt{\mathrm{trace}(\Theta^T\Theta)} = \sqrt{\sum_{j=1}^m \sigma_j^2(\Theta)}$. We refer the reader to Horn and Johnson (13) 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 $\Theta^* \in \mathbb{R}^{k \times p}$ is equivalent to requiring that the rows (or columns) of Θ^* lie in some r-dimensional subspace of $\mathbb{R}^p$ (or $\mathbb{R}^k$ respectively). 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 $Z_j \in \mathbb{R}^p$ to multi-dimensional output vectors $Y_j \in \mathbb{R}^k$. More specifically, let us consider the linear model, specified by a matrix $\Theta^* \in \mathbb{R}^{k \times p}$, of the form

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

where $\{W_a\}_{a=1}^n$ is an i.i.d. sequence of k-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, neural response modeling, and analysis of financial data. This model and closely related ones also arise in the problem of collaborative filtering (5), 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.

As a second (not unrelated) example, we consider the problem of system identification in vector autoregressive processes (see the book (14) for a detailed background). A *vector autoregressive* (VAR) process in p-dimensions is a stochastic process $\{Z_t\}_{t=1}^\infty$ specified by an initialization $Z_1 \in \mathbb{R}^p$, followed by the recursion

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

In this recursion, the sequence $\{W_t\}_{t=1}^\infty$ consists of i.i.d. samples of innovations noise. We assume that each vector $W_t \in \mathbb{R}^p$ is zero-mean with covariance $\nu^2 I$, 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

$$\Sigma = \Theta^* \Sigma (\Theta^*)^T + \nu^2 I. \qquad (3)$$

The goal of system identification in a VAR process is to estimate the unknown matrix $\Theta^* \in \mathbb{R}^{p \times p}$ 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 p 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, subspace tracking models in signal processing, and motion models in computer vision.

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 (2; 3), and Recht et al. (9) suggested and studied its extension to low-rank matrices. In their set-up, one observes trace inner products of the form $\langle\langle X_i,\ \Theta^* \rangle\rangle = \text{trace}(X_i^T \Theta^*)$, where $X_i \in \mathbb{R}^{k \times p}$ is a random matrix (for instance, filled with standard normal $N(0,1)$ 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 database at two different time instants), and matrix denoising.

A fourth example that can also be treated within our framework is the *matrix completion model*, in which each observation matrix takes the form $X_i = e_{a(i)} e_{b(i)}^T$, so that X_i is non-zero except at a randomly chosen pair $(a(i), b(i))$ of row/column indices. This problem has been studied by several authors in recent work (e.g., (5; 10; 8; 11)).

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 $A, B \in \mathbb{R}^{k \times p}$, recall the Frobenius or trace inner product $\langle\langle A,\ B \rangle\rangle := \text{trace}(BA^T)$. 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

$$\vec{y} = \begin{bmatrix} y_1 & \ldots & y_N \end{bmatrix}^T \in \mathbb{R}^N,$$

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 $\mathfrak{X} : \mathbb{R}^{k \times p} \to \mathbb{R}^N$ via $\big[\mathfrak{X}(\Theta)\big]_i = \langle\langle X_i,\ \Theta \rangle\rangle$. With this notation, the observation model (4) can be re-written as

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

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 $\Theta \in \mathbb{R}^{k \times p}$ is given by $|||\Theta|||_1 = \sum_{j=1}^{\min\{k,p\}} \sigma_j(\Theta)$, corresponding to the sum of its singular values. Given a collection of observations $(y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{k \times p}$, for $i = 1, \ldots, N$ from the observation model (4), we consider estimating the unknown Θ^* by solving the following optimization problem

$$\widehat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^{k \times p}} \Big\{ \frac{1}{2N} \|\vec{y} - \mathfrak{X}(\Theta)\|_2^2 + \lambda_N |||\Theta|||_1 \Big\}, \tag{6}$$

where $\lambda_N > 0$ is a regularization parameter. Note that the optimization problem (6) can be viewed as the analog of the Lasso estimator, tailored to low-rank matrices as opposed to sparse vectors. An important property of the optimization problem (6) is that it can be solved in time polynomial in the sample size N and the matrix dimensions k and p. Indeed, the optimization problem (6) is an instance of a *semidefinite program*, a class of convex optimization problems that can be solved efficiently by various polynomial-time algorithms. For instance, when the problem parameters are small, interior point methods are a classical method that can be employed for solving the semidefinite programs. However, as we discuss in Section 4, there are a variety of other methods tailored to solving our specific M-estimation procedure that lend themselves to solving larger-scale problems.

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 that guarantee that the estimate $\widehat{\Theta}$ is close in Frobenius norm to the unknown matrix Θ^*.

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 3.2 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 (12), since it amounts to guaranteeing that the quadratic loss function in the SDP (6) is strictly convex over a restricted set of directions. Letting $\mathcal{C} \subseteq \mathbb{R}^{k \times p}$ 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

$$\frac{1}{2N}\|\mathfrak{X}(\Delta)\|_2^2 \geq \kappa(\mathfrak{X}) \, |||\Delta|||_F^2 \qquad \text{for all } \Delta \in \mathcal{C}. \quad (7)$$

We note that analogous conditions have been used to establish error bounds in the context of sparse linear regression (1), in which case the set $\mathcal{C}$ corresponded to certain subsets of sparse vectors. Of course, the definition (7) hinges on the choice of the restricted set $\mathcal{C}$. In order to define the set, we require some additional notation. For any matrix $\Theta \in \mathbb{R}^{k \times p}$, we let $\text{row}(\Theta) \subseteq \mathbb{R}^p$ and $\text{col}(\Theta) \subseteq \mathbb{R}^k$ denote its row space and column space, respectively. For a given positive integer $r \leq \min\{k, p\}$, any r-dimensional subspace of $\mathbb{R}^k$ can be represented by some orthogonal matrix $U \in \mathbb{R}^{k \times r}$ (i.e., that satisfies $U^T U = I_{r \times r}$. In a similar fashion, any r-dimensional subspace of $\mathbb{R}^p$ can be represented by an orthogonal matrix $V \in \mathbb{R}^{p \times r}$. For any fixed pair of such matrices (U, V), we may define $\mathcal{M}(U, V)$ as the set of Θ such that $\text{row}(\Theta) \subset V$ and $\text{col}(\Theta) \subset U$ and $\mathcal{M}^{\perp}(U, V)$ as the set of Θ such that $\text{row}(\Theta) \perp V$ and $\text{col}(\Theta) \perp V$. Finally, we let $\Pi_{\mathcal{M}(U,V)}$ and $\Pi_{\mathcal{M}^{\perp}(U,V)}$ denote the (respective) projection operators onto these subspaces. When the subspaces (U, V) are clear from context, we use the shorthand notation $\Delta'' = \Pi_{\mathcal{M}^{\perp}(U,V)}(\Delta)$ and $\Delta' = \Delta - \Delta''$. Finally, for any positive integer $r \leq \min\{k, p\}$, we let (U^r, V^r) denote the subspace pair defined by the top r left and right singular vectors of Θ^*. For a given integer r and tolerance $\delta > 0$, we then define a subset of matrices as follows:

$$\mathcal{C}(r; \delta) := \Big\{ \Delta \in \mathbb{R}^{k \times p} \mid |||\Delta|||_F \geq \delta, \; |||\Delta''|||_1 \leq 3|||\Delta'|||_1 + 4|||\Pi_{\mathcal{M}^{\perp}(U^r, V^r)}(\Theta^*)|||_1 \Big\}. \quad (8)$$

The next ingredient is the choice of the regularization parameter λ_N used in solving the SDP (6). Our theory specifies a choice for this quantity in terms of the adjoint of the operator $\mathfrak{X}$—namely, the operator $\mathfrak{X}^* : \mathbb{R}^N \to \mathbb{R}^{k \times p}$ defined by

$$\mathfrak{X}^*(\vec{\varepsilon}) := \sum_{i=1}^N \varepsilon_i X_i. \quad (9)$$

With this notation, we now state a deterministic result, analogous to the main result in our past work (12), which specifies two conditions—namely, an RSC condition and a choice of the regularizer—that suffice to guarantee that any solution of the SDP (6) fall within a certain radius.

Theorem 1. *Suppose 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 $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\text{op}}/N$. Then any solution $\widehat{\Theta}$ to the semidefinite program (6) satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\Big\{\delta, \; \frac{32\lambda_N \sqrt{r}}{\kappa(\mathfrak{X})}, \; \Big[\frac{16 \, \lambda_N \, |||\Pi_{\mathcal{M}^{\perp}(U^r, V^r)}(\Theta^*)|||_1}{\kappa(\mathfrak{X})}\Big]^{1/2}\Big\}. \quad (10)$$

Apart from the tolerance parameter δ, the two main terms in the bound (10) 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*, in which the projection onto the set $\mathcal{M}^{\perp}(U^r, V^r)$ describes the gap between the true matrix Θ^* and the rank r approximation.

Let us begin by illustrating the consequences of Theorem 1 when the true matrix Θ^* has exactly rank r, in which case there is a very natural choice of the subspaces represented by U and V. In particular, we form U from the r non-zero left singular vectors of Θ^*, and V from its r non-zero right singular vectors. Note that this choice of (U, V) ensures that $\Pi_{\mathcal{M}^{\perp}(U,V)}(\Theta^*) = 0$. For technical reasons, 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 Θ^* has rank r, and $\mathfrak{X}$ satisfies RSC with respect to $\mathcal{C}(r; 0)$. Then as long as $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\text{op}}/N$, any optimal solution $\widehat{\Theta}$ to the SDP (6) satisfies the bound*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \frac{32\sqrt{r} \, \lambda_N}{\kappa(\mathfrak{X})}. \quad (11)$$

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 $\{\sigma_i(\Theta^*)\}_{i=1}^{\min\{k,p\}}$ decays quickly enough. In particular, for a parameter $q \in [0,1]$ and a positive radius R_q, we define the set

$$\mathbb{B}_q(R_q) := \big\{\Theta \in \mathbb{R}^{k\times p} \mid \sum_{i=1}^{\min\{k,p\}} |\sigma_i(\Theta)|^q \leq R_q\big\}.$$

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 $\Theta^* \in \mathbb{B}_q(R_q)$, the regularization parameter is lower bounded as $\lambda_N \geq 2|||\mathfrak{X}^*(\vec{\varepsilon})|||_{\rm op}/N$, 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 (6) satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F \leq \max\Big\{\delta,\ 32\sqrt{R_q}\Big(\frac{\lambda_N}{\kappa(\mathfrak{X})}\Big)^{1-q/2}\Big\}. \quad (12)$$

Note that the error bound (12) reduces to the exact low rank case (11) when $q = 0$, and $\delta = 0$. The quantity $\lambda_N^{-q}R_q$ acts as the "effective rank" in this setting. 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. 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. 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. Recall that we observe pairs $(Y_i, Z_i) \in \mathbb{R}^k \times \mathbb{R}^p$ linked by the linear model $Y_i = \Theta^* Z_i + W_i$, where $W_i \sim N(0, \nu^2 I_{k\times k})$ 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 p-dimensional covariance matrix $\Sigma \succ 0$. Recalling that $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ denote the maximum and minimum eigenvalues respectively, we have:

Corollary 3 (Low-rank multivariate regression)**.** *Consider the random design multivariate regression model where $\Theta^* \in \mathbb{B}_q(R_q)$. There are universal constants $\{c_i, i = 1,2,3\}$ such that if we solve the SDP (6) with regularization parameter $\lambda_N = 10\nu\sqrt{\sigma_{\max}(\Sigma)}\sqrt{\frac{(k+p)}{n}}$, we have*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1\left(\frac{\nu^2\sigma_{\max}(\Sigma)}{\sigma^2_{\min}(\Sigma)}\right)^{1-q/2} R_q\left(\frac{k+p}{n}\right)^{1-q/2} \quad (13)$$

with probability greater than $1 - c_2\exp(-c_3(k+p))$.

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

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

The scaling in this error bound is easily interpretable: naturally, the squared error is proportional to the noise variance ν^2, and the quantity $r(k+p)$ counts the number of degrees of freedom of a $k \times p$ matrix with rank r. Note that if we did not impose any constraints on Θ^*, then since a $k \times p$ matrix has a total of kp free parameters, we would expect at best to obtain rates of the order $|||\widehat{\Theta} - \Theta^*|||_F^2 = \Omega(\frac{\nu^2\, k\, p}{n})$. Note that when Θ^* is low rank—in particular, when $r \ll \min\{k,p\}$—then the nuclear norm estimator achieves substantially faster rates. Finally, we note that as stated, the result requires that $\min\{k,p\}$ 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 (k,p), the same statement (with different constants) holds with $k+p$ replaced by $\log n$.

Next we turn to the case of estimating the system matrix Θ^* of an autoregressive (AR) model.

Corollary 4 (Autoregressive models)**.** *Suppose that we are given n samples $\{Z_t\}_{t=1}^n$ from a p-dimensional autoregressive process (2) that is stationary, based on a system matrix that is stable ($|||\Theta^*|||_{\rm op} \leq \gamma < 1$), and approximately low-rank ($\Theta^* \in \mathbb{B}_q(R_q)$). Then there are universal constants $\{c_i, i = 1, 2, 3\}$ such that if we solve the SDP (6) with regularization parameter $\lambda_N = \frac{80\,|||\Sigma|||_{\rm op}}{1-\gamma}\sqrt{\frac{p}{n}}$, then any solution $\widehat{\Theta}$ satisfies*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq c_1 \left[\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)\,(1-\gamma)}\right]^{2-q} R_q \left(\frac{p}{n}\right)^{1-q/2} \tag{14}$$

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

Remarks: Like Corollary 3, the result as stated requires that p tend to infinity, but the same bounds hold with p replaced by $\log n$, yielding results suitable for classical (fixed dimension) scaling. Second, the factor $(p/n)^{1-q/2}$, like the analogous term[1] 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 rp/n. In contrast to the case of multivariate regression, the error bound (14) also depends on the upper bound $|||\Theta^*|||_{\rm op} = \gamma < 1$ 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. The dependence of the sampling in the AR model introduces 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. The following result applies to the setting in which the observation matrices $\{X_i\}_{i=1}^N$ are drawn i.i.d., with standard $N(0, 1)$ elements. We assume that the observation noise vector $\vec{\varepsilon} \in \mathbb{R}^N$ satisfies the bound $\|\vec{\varepsilon}\|_2 \leq 2\nu\sqrt{N}$ for some constant ν, an 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 (one example being Gaussian noise $N(0, \nu^2)$).

Corollary 5 (Compressed sensing recovery)**.** *Suppose that $\Theta^* \in \mathbb{B}_q(R_q)$, and that the sample size is lower bounded as $N > 4\max(k, p)\,(100\,R_q)^{2/(2-q)}$. Then any solution $\widehat{\Theta}$ to the SDP (6) satisfies the bound*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \leq 256\,\nu^{2-q}\,R_q\left[\sqrt{\frac{k}{N}} + \sqrt{\frac{p}{N}}\right]^{2-q} \tag{15}$$

with probability greater than $1 - c_1\exp(-c_2(k + p))$.

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. *Under the stated conditions, the random operator $\mathfrak{X}$ satisfies*

$$\frac{\|\mathfrak{X}(\Theta)\|_2}{\sqrt{N}} \geq \frac{1}{4}|||\Theta|||_F - \left(\sqrt{\frac{k}{N}} + \sqrt{\frac{p}{N}}\right)|||\Theta|||_1 \quad \text{for all } \Theta \in \mathbb{R}^{k\times p} \tag{16}$$

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

Proposition 1 also implies an interesting property of the null space of the operator $\mathfrak{X}$; one that 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

$$\min_{\Theta\in\mathbb{R}^{k\times p}} |||\Theta|||_1 \quad \text{s.t. } \langle\langle X_i,\ \Theta\rangle\rangle = y_i \quad \forall i = 1, \ldots, N, \tag{17}$$

a recovery procedure that was studied by Recht et al. (9). Proposition 1 allows us to obtain a sharp result on recovery using this method:

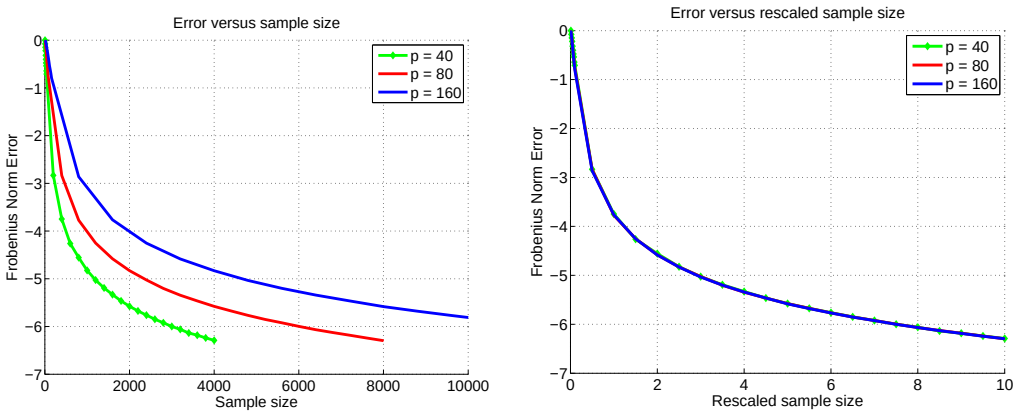
Corollary 6. *Suppose that Θ^* has rank r, and that we are given $N > 40^2 r(k + p)$ noiseless samples. Then with probability at least $1 - 2\exp(-N/32)$, the SDP (17) recovers the matrix Θ^* exactly.*

This result removes some extra logarithmic factors that were included in the earlier work (9), and provides the appropriate analog to compressed sensing results for sparse vectors (2). 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 $k + p$.

4. 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

[1] The term in Corollary 3 has a factor $k + p$, since the matrix in that case could be non-square in general.



(a) (b)

Figure 1. Results of applying the SDP (6) 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 $p \in \{40, 80, 160\}$, all with rank $r = 10$. (b) Plots of the same Frobenius error versus the rescaled sample size $N/(rp)$. Consistent with theory, all three plots are now extremely well-aligned.

behavior of the SDP-based M-estimator (6) in practice. In all cases, we solved the convex program (6) by using our own implementation in MATLAB of an accelerated gradient descent method which adapts a non-smooth convex optimization procedure (15) to the nuclear-norm (16). We chose the regularization parameter λ_N in the manner suggested by our theoretical results; in doing so, we assumed knowledge of quantities such as the noise variance ν^2. (In practice, one would have to estimate such quantities from the data using standard methods.)

We report simulation results for problems of 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 $\Theta^* \in \mathbb{R}^{p \times p}$, where $p \in \{40, 80, 160\}$ for the first two examples, and $p \in \{20, 40, 80\}$ for the compressed sensing example. In all cases, we considered the setting of exact low rank constraints, with $\text{rank}(\Theta^*) = r = 10$, and we generated Θ^* by choosing the subspaces of its left and right singular vectors uniformly at random from the Grassman manifold. The observation or innovations noise had variance $\nu^2 = 1$ in each case. 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, p), we solved the SDP for various sample sizes N.

Figure 1 shows results for a multivariate regression model with the covariates chosen randomly from a $N(0, I)$ distribution. 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 p is increased. We note that panel (b) shows the exact same set of simulation results, but now with the Frobenius error plotted versus the rescaled sample size $\widetilde{N} := N/(rp)$. 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, the figures show that $N/(rp)$ acts as the effective sample size in this high-dimensional setting.

Figure 2 shows similar results for the autoregressive model. The figure plots the Frobenius error versus the rescaled sample size $N/(rp)$; as predicted by Corollary 4, the errors for different matrix sizes p 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. 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/rp, as predicted by Corollary 5.

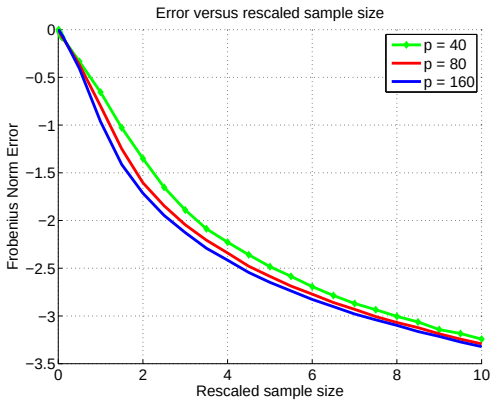


Figure 2. Results of applying the SDP (6) to estimating the system matrix of a vector autoregressive process. Plot of the Frobenius error versus the rescaled sample size $N/(rp)$. Consistent with theory, all three plots are now reasonably well-aligned.

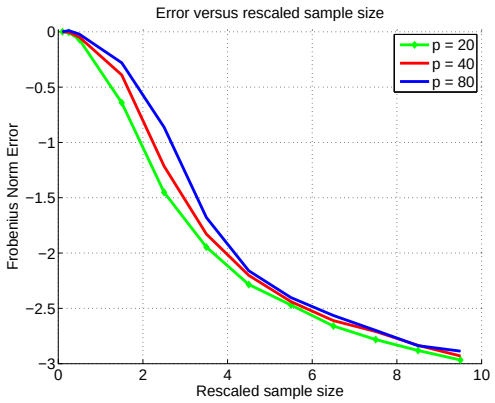


Figure 3. Results of applying the SDP (6) to recovering a low-rank matrix on the basis of random projections (compressed sensing model). Plot of the same Frobenius error versus the rescaled sample size $N/(rp)$. Consistent with theory, all three plots are now reasonably well-aligned.

5. Discussion

In this paper, we provided a detailed analysis of the nuclear norm relaxation for a general class of noisy observation models, and obtained non-asymptotic error bounds on the Frobenius norm valid under high-dimensional scaling. Our results are applicable to both exactly and approximately low-rank matrices. Exploiting a deterministic result that leverages our past work (12), we showed concrete and easily interpretable rates for various specific models, including low-rank multivariate regression, estimation of autoregressive processes, and matrix recovery from random projections. It is worth noting that our theory can also be applied to noisy matrix completion, yielding analogous rates to those reported here. Lastly, our simulation results showed very close agreement with the predictions from our theory.

References

[1] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *Annals of Statistics*, vol. 3, pp. 38–60, 2009.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[3] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Info Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.

[4] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford, 2002.

[5] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *NIPS*, Vancouver, Canada, December 2004.

[6] J. Abernethy, F. Bach, T. Evgeniou, and J. Stein, "Low-rank matrix factorization with attributes," Tech. Rep., September 2006.

[7] F. Bach, "Consistency of trace norm minimization," *Journal of Machine Learning Research*, vol. 9, pp. 1019–1048, June 2008.

[8] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," 2009, preprint available at http://arxiv.org/abs/0906.2027v1.

[9] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, 2007.

[10] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *CoRR*, vol. abs/0805.4471, 2008.

[11] R. Mazumber, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," Stanford, Tech. Rep., July 2009.

[12] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *NIPS*, December 2009.

[13] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[14] H. Lütkepolhl, *New introduction to multiple time series analysis*. New York: Springer, 2006.

[15] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE, Universit'e catholique de Louvain, Tech. Rep. 2007/76, 2007.

[16] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *ICML*, 2009.