Sparse and Low-Rank Matrix Decompositions ⋆

Venkat Chandrasekaran * **Sujay Sanghavi** ** **Pablo A. Parrilo** *
Alan S. Willsky *

* *Laboratory for Information and Decision Systems*
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
(email: venkatc@mit.edu; parrilo@mit.edu; willsky@mit.edu)
** *School of Electrical and Computer Engineering*
Purdue University
West Lafayette, IN 47907 USA
(e-mail: sanghavi@ecn.purdue.edu)

Abstract: Suppose we are given a matrix that is formed by adding an unknown sparse matrix to an unknown low-rank matrix. Our goal is to decompose the given matrix into its sparse and low-rank components. Such a problem arises in a number of applications in model and system identification, but obtaining an exact solution is NP-hard in general. In this paper we consider a *convex optimization* formulation to splitting the specified matrix into its components; in fact our approach reduces to solving a semidefinite program. We provide sufficient conditions that guarantee *exact recovery* of the components by solving the semidefinite program. We also show that when the sparse and low-rank matrices are drawn from certain natural random ensembles, these sufficient conditions are satisfied with high probability. We conclude with simulation results on synthetic matrix decomposition problems.

1. INTRODUCTION

In many engineering applications one encounters complex systems and models that are often composed of multiple simpler systems and models. In order to better understand the behavior of the complex system, a natural approach is to break it down into simpler components. In this paper we consider matrix representations of systems and statistical models in which our matrices are formed by adding together *sparse* and *low-rank* matrices. The object of this paper is to propose a *tractable* solution for recovering the sparse and low-rank components, and to analyze when our approach recovers these components *exactly*.

The low-rank and sparse matrices have different interpretations based on the problem at hand. In a statistical model selection setting, the sparse matrix can correspond to a sparse Gaussian graphical model (Lauritzen [1956]), while the low-rank matrix can be attributed to the effect due to marginalization of latent variables. Being able to decompose a matrix into simpler sparse and low-rank components can have important implications for the development of efficient estimation algorithms. In a system identification setting, the low-rank (Hankel) matrix corresponds to a low-order LTI system, and the sparse (Hankel) matrix can capture an LTI system with a sparse impulse response. Such a decomposition can be used to provide a simpler, more efficient description of a complex system in terms of its simpler components.

⋆ This work was supported in part by a MURI funded through AFOSR Grant FA9550-06-1-0324, and by a MURI funded through ARO Grant W911NF-06-1-0076.

For a low-rank matrix with entries perturbed by Gaussian noise, one can recover an estimate of the low-rank matrix by using techniques based on the singular value decomposition (SVD). Such an approach might involve some form of thresholding of the singular values. However, the problem that we consider is rather different in that a low-rank matrix is "perturbed" by a *sparse* matrix. The sparse matrix can in general have entries of *arbitrary* magnitude. Therefore, SVD-based methods are not directly applicable in order to solve such problems. Solving this problem is NP-hard in general, but we consider approaches employing recently well-studied convex relaxations.

In formulating our convex optimization problem, we use the ℓ_1 norm as a surrogate for counting the number of non-zero entries based on the numerous results that demonstrate its effectiveness at recovering sparse solutions (see for example Donoho [2006a]). We also use the nuclear norm, which is the sum of the singular values of a matrix, to replace the rank of the matrix (Fazel [2002]). This is a generalization of the previously studied trace-heuristic used in various control problems (Mesbahi et al. [1997]). Recent work has also demonstrated that the nuclear norm is an effective surrogate in rank minimization problems (Recht et al. [2007], Candes and Recht [2008]).

Suppose that we are given a matrix $C = A^* + B^*$ with A^* an unknown sparse matrix and B^* an unknown low-rank matrix. We seek to recover the matrices A^* and B^*. We consider the following convex optimization problem:

$$\min_{A,B} \quad \gamma\|A\|_{\ell_1} + \|B\|_* \qquad s.t. \quad A + B = C.$$

Here, γ is a parameter that trades off between the low-rank and sparse components. Clearly, one cannot expect to recover from $C = A^* + B^*$ arbitrary sparse and low-rank matrices A^*, B^* as such a problem is ill-posed in general. We provide a set of sufficient conditions under which the solution of the convex optimization problem above yields the *exact* A^*, B^*. These conditions essentially require that the space spanned by the singular vectors of B^* is "incoherent" with the standard basis, while the sparse matrix A^* does not contain too many non-zero elements in each row/column. We also show that these sufficient conditions are satisfied with high probability when A^*, B^* are drawn from certain natural random ensembles of sparse and low-rank matrices. Our convex program can in fact be cast as a semidefinite program (SDP) (Vandenberghe and Boyd [1996]), which can be solved efficiently in polynomial-time.

Relation to previous work The ℓ_1 norm has been well-studied as a heuristic for finding sparse representations of signals based on overcomplete dictionaries (Donoho and Elad [2003]). Building on these results, *compressed sensing* (Candes et al. [2006], Donoho [2006b]) was introduced as a framework for recovering a sparse signal based on very few measurements. Analogous to the compressed sensing setting, a number of recent papers (Recht et al. [2007], Candes and Recht [2008]) have addressed the problem of recovering a low-rank *matrix* given very few linear functionals of the matrix (e.g., a subset of the entries of the matrix). Our work is closer in spirit to that of Donoho and Elad [2003], and can be viewed as a method to find the "simplest" representation of a matrix given an "overcomplete dictionary" of sparse matrix atoms and low-rank matrix atoms.

Paper outline In Section 2, we introduce a problem from system identification and one from statistical model selection to demonstrate the potential applicability of our work. In Section 3 we present the formal problem statement and discuss the kinds of sparse and low-rank matrices that one can hope to recover. Section 4 introduces our optimization formulation and provides a set of sufficient conditions for exact recovery based on subgradient optimality conditions. In Section 5 we provide an even simpler set of sufficient conditions for exact recovery, and apply these conditions to demonstrate that exact recovery is possible under a certain probabilistic regime with high probability. We describe the results of simulations of our approach applied to synthetic matrix decomposition problems in Section 6, and conclude with a brief discussion in Section 7.

Our analysis can be extended to provide an uncertainty principle that quantifies a notion of rank-sparsity incoherence for arbitrary matrices. Due to space constraints we do not discuss this point in detail, and we also do not provide any proofs of our results in this paper. We defer these to a longer report.

2. STYLIZED APPLICATIONS

Our work has important applications in model selection. Given a joint covariance matrix $\Sigma_{(o\ h)}$ on a collection of observed variables o and hidden variables h, the marginal covariance of the observed variables is simply the corresponding submatrix Σ_o. In many applications, however, Gaussian models are parameterized in terms of the inverse covariance matrix $J_{(o\ h)} = \Sigma^{-1}_{(o\ h)}$ (also known as the precision or concentration or information matrix). Such a parameterization reveals the connection to sparse Gauss-Markov models in which the information matrix is sparse according to the underlying graph on which the Gaussian variables obey the Markov property (Lauritzen [1956]). The *marginal information matrix* of the observed variables Σ_o^{-1} is given by a Schur complement relation:

$$\hat{J}_o = J_o - J_{o,h} J_h^{-1} J_{h,o}, \tag{1}$$

where J_o is typically sparse due to sparsity in the graphical structure, while $J_{o,h} J_h^{-1} J_{h,o}$ has rank equal to the number of latent variables h. Thus, decomposing $\hat{J}_o$ into these components reveals the graphical structure in the observed variables as well as the effect due to (and the *number* of) the unobserved latent variables.

A similar problem can be posed in the system identification setting in which the input-output relation of an LTI system is described by a matrix H as follows:

$$H = H_s + H_{lr}.$$

Here, H_{lr} is a low-rank Hankel matrix corresponding to a low-order system. The matrix H_s is a sparse Hankel matrix and describes an LTI system with a sparse impulse response. Decomposing H into these components could potentially provide a simpler description of the system rather than considering the system matrix H by itself.

We note that our work also has applications in problems from computer science such as those involving matrix rigidity (Valiant [1977]).

3. PRELIMINARIES AND PROBLEM STATEMENT

3.1 Notation

We briefly describe the notation used in this paper. We consider matrices that live in the space [1] $\mathbb{R}^{n \times n}$. The support of a matrix M (the locations of the non-zero entries in M) is denoted $\text{supp}(M) \subseteq \{1, \cdots, n\} \times \{1, \cdots, n\}$. The rank of a matrix M is denoted $\text{rank}(M)$. The following matrix norms will be employed throughout this paper:

- $\|\cdot\|$ refers to the spectral norm, or the largest singular value.
- $\|\cdot\|_{\ell_\infty}$ refers to the absolute element-wise maximum entry of a matrix (not the operator ℓ_∞ norm).
- $\|\cdot\|_*$ refers to the nuclear norm, or the sum of the singular values.
- $\|\cdot\|_{\ell_1}$ refers to the element-wise sum of absolute values (not the operator ℓ_1 norm).
- $\|\cdot\|_F$ refers to the Frobenius norm.

[1] All our results extend to the rectangular case, but we stick with the square case to avoid cluttered notation. If the matrices belong to $\mathbb{R}^{n_1 \times n_2}$, then our results can be extended by simply setting $n = \max(n_1, n_2)$.

We will not have occasion to use the operator ℓ_1 and ℓ_∞ matrix norms. Therefore, there should be no ambiguity with the above notation. We also use the ℓ_2 vector norm, and will denote it by $\|\cdot\|_2$.

3.2 Problem Statement and Rank-Sparsity incoherence

Main Problem Given a matrix $C = A^* + B^*$ where A^* is sparse and B^* is low-rank, recover the components A^* and B^*.

Let $\text{supp}(A^*)$ be the support of A^*, and let B^* be a rank-k matrix with singular value decomposition (SVD) $B^* = U\Sigma V'$. Here, $U \in \mathbb{R}^{n\times k}, \Sigma \in \mathbb{R}^{k\times k}$, and $V \in \mathbb{R}^{n\times k}$.

It is apparent that our main problem is ill-posed; we describe two such situations below. They provide insight into the kind of additional conditions that must be imposed on A^*, B^* in order to guarantee exact recovery. First, suppose that B^* consists of a 1 in the top-left location and 0 everywhere else. While such a B^* is low-rank, it is also very sparse. One could not reasonably expect to recover such a B^* from C, since a valid sparse-plus-low-rank decomposition $(\hat{A}, \hat{B})$ is $\hat{A} = A^* + B^*, \hat{B} = 0$. Therefore, we need an appropriate notion of low-rank that ensures that B^* is not too sparse. As will be seen later in the paper, one natural way to accomplish this is to impose conditions that require the space spanned by the singular vectors U and V (i.e., the row and column spaces of B^*) to be "incoherent" with the standard basis. Similar conditions have been discussed in recent work on the matrix completion problem (Candes and Recht [2008]). Second, suppose that the first column of A^* is non-zero, while all other columns are 0. Thus, A^* is relatively sparse, but it is also low-rank (it has rank 1). In such a scenario a reasonable decomposition $(\hat{A}, \hat{B})$ might be $\hat{A} = 0, \hat{B} = A^* + B^*$ (here $\text{rank}(\hat{B}) \leq \text{rank}(B^*) + 1$). Consequently, we consider sparse matrices that have "bounded degree", i.e., each row/column does not have too many non-zero entries, to avoid such identifiability issues.

4. CONVEX OPTIMIZATION FORMULATION

In this section, we motivate the convex optimization problem that was stated in the introduction. The general problem of recovering A^* and B^* from C is NP-hard, and a typical solution might involve a search with combinatorial complexity. However, a number of heuristics have been developed for approximating the functions $|\text{supp}(\cdot)|$ and $\text{rank}(\cdot)$. In particular the $\|\cdot\|_{\ell_1}$ norm has proved to be surprisingly effective as a surrogate for $|\text{supp}(\cdot)|$ and there is a vast literature describing conditions under which one can recover a sparse signal *exactly* using the ℓ_1 heuristic (Donoho [2006a]). Note that $\|\cdot\|_{\ell_1}$ is a *convex* function, unlike $|\text{supp}(\cdot)|$. More recently, a growing body of work has advocated the use of the nuclear norm $\|\cdot\|_*$ in place of $\text{rank}(\cdot)$ (Fazel [2002], Recht et al. [2007], Candes and Recht [2008]). This is a generalization of the trace-heuristic previously employed as a surrogate for the rank of a positive-semidefinite matrix (Mesbahi et al. [1997]). Notice that the nuclear norm is a convex function, and is analogous to an "ℓ_1 norm" applied to the singular values. A variety of results provide conditions under which one can recover a low-rank matrix *exactly* using the nuclear norm heuristic. We consider the following optimization program to split a given matrix C into its sparse and low-rank components:

$$\begin{aligned} \min_{A,B} \quad & \gamma\|A\|_{\ell_1} + \|B\|_* \\ s.t. \quad & A + B = C. \end{aligned} \tag{2}$$

Here, γ is a regularization parameter that provides a trade-off between the two components. In situations where C contains additional noise, one can relax the equality constraints. In this paper, we focus on the problem with no additional noise, i.e., the case with equality constraints. We note that (2) can be cast as an SDP (see Appendix A) and can be solved efficiently with complexity *polynomial* in the size of the input using off-the-shelf SDP solvers.

4.1 Optimality conditions

Before describing the subgradient-based optimality conditions for (2), we introduce certain spaces associated with the matrics A^* and B^*. We consider the space of all matrices that have the same support as A^*:

$$\Omega = \{M \in \mathbb{R}^{n\times n} |\ \text{supp}(M) \subseteq \text{supp}(A^*)\}. \tag{3}$$

The projection of a matrix onto the space Ω is denoted by P_Ω, which simply sets to zero those entries with support not in $\text{supp}(A^*)$. Let Ω^c denote the orthogonal space to Ω, i.e., the space of those matrices with support in $\text{supp}(A^*)^c$. The projection onto Ω^c is denoted P_{Ω^c}.

The SVD of $B^* = U\Sigma V'$, where as described previously, $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{n\times k}$. Consider the following tangent space at B^* to the manifold of rank-k matrices:

$$T = \{UX' + YV' |\ X \in \mathbb{R}^{n\times k}, Y \in \mathbb{R}^{n\times k}\}. \tag{4}$$

The space T is generated by the span of all matrices that either have the same column space as B^* or the same row space as B^*. The projection of a matrix onto the space T is given by P_T:

$$P_T(M) = P_U M + M P_V - P_U M P_V,$$

where $P_U = UU'$ and $P_V = VV'$ denote the projection onto the spaces spanned by U and V respectively. Let $T^\perp$ denote the orthogonal space to T, i.e., the span of all matrices with column space orthogonal to the column space of B^* *and* row space orthogonal to the row space of B^*. We denote the projection onto this space by $P_{T^\perp}$:

$$P_{T^\perp}(M) = (I_n - P_U)M(I_n - P_V).$$

The following proposition gives a set of sufficient conditions for (A^*, B^*) to be the *unique* optimum of (2). Let $\text{sign}(\cdot)$ denote the element-wise signum function of a matrix.

Proposition 1. Suppose that $C = A^* + B^*$. Then, (A^*, B^*) is the unique *optimizer of (2) if the following conditions are satisfied:*

(1) Only the 0 *matrix belongs to both Ω and T, i.e., $\Omega \cap T = \{0\}$.*
(2) There exists a dual $Q \in \mathbb{R}^{n\times n}$ such that
  (a) $P_T(Q) = UV'$
  (b) $P_\Omega(Q) = \gamma\text{sign}(A^*)$
  (c) $\|P_{T^\perp}(Q)\| < 1$
  (d) $\|P_{\Omega^c}(Q)\|_{\ell_\infty} < \gamma$

Figure 1 provides a visual representation of these conditions. In particular, we see that the spaces Ω and T only

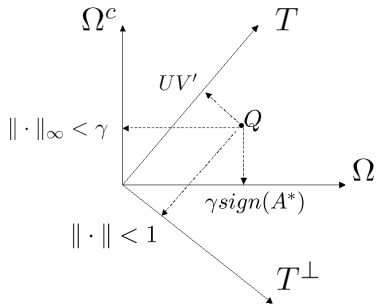

Fig. 1. Geometric view of optimality conditions: Existence of a dual Q.

intersect at 0. One can also intuitively see that guaranteeing the existence of a dual Q with the requisite conditions is perhaps easier if the "angle" between Ω and T is greater.

Applying optimality conditions from convex analysis (Bertsekas et al. [2003]), one can check that (A^*, B^*) is *an* optimum of (2) if there exists a dual $Q \in \mathbb{R}^{n\times n}$ such that

$$Q \in \gamma\partial\|A^*\|_1 \text{ and } Q \in \partial\|B^*\|_*. \tag{5}$$

For $Q \in \gamma\partial\|A^*\|_1$, we must have that

$$P_\Omega(Q) = \gamma \text{sign}(A^*), \quad \|P_{\Omega^c}(Q)\|_\infty \leq \gamma. \tag{6}$$

For $Q \in \partial\|B^*\|_*$, we must have that (Watson [1992])

$$P_T(Q) = UV', \quad \|P_{T^\perp}(Q)\| \leq 1. \tag{7}$$

As noted above, (5) only provides a set of conditions for (A^*, B^*) to be an optimum of (2). The conditions in Proposition (1) involve some tightening of these subgradient conditions (5), (6), and (7) in order to guarantee uniqueness.

5. SUFFICIENT CONDITIONS FOR EXACT RECOVERY

5.1 Simple sufficient conditions

We describe a set of simpler sufficient conditions that guarantee the existence of a dual Q in Proposition 1. Consider the following two quantities:

$$\xi_T = \max_{M\in T, \|M\|\leq 1} \|M\|_\infty \tag{8}$$

$$\xi_\Omega = \max_{M\in \Omega, \|M\|_\infty \leq 1} \|M\| \tag{9}$$

If ξ_T is small, then we have that elements of T are "diffuse", i.e., a matrix in T cannot have very large entries.[2] Similarly, if ξ_Ω is small then we have that the *spectrum* of an element of Ω is "diffuse", i.e., has small singular values.

The following proposition provides a simpler set of sufficient conditions than those provided by Proposition 1.

Proposition 2. If $\xi_T\xi_\Omega < \frac{1}{8}$, then $\Omega \cap T = \{0\}$ and there exists a dual Q such that the second set of conditions $2(a) - 2(d)$ in Proposition (1) is satisfied for

$$\gamma \in \left(\frac{\xi_T}{1 - 6\xi_T\xi_\Omega}, \frac{1 - 4\xi_T\xi_\Omega}{\xi_\Omega}\right). \tag{10}$$

Consequently, $\xi_T\xi_\Omega < \frac{1}{8}$ implies that (A^*, B^*) is the unique *optimizer of (2).*

[2] Specifically, unit-spectral-norm matrices in T cannot have very large entries.

Observe that if $T = \mathbb{R}^{n\times n}$, then $\xi_T = 1$. Similarly, if $\Omega = \mathbb{R}^{n\times n}$, then $\xi_\Omega = n$. Therefore, we clearly need to use special structure in T and in Ω in order to identify classes of matrices in which $\xi_T\xi_\Omega < \frac{1}{8}$.

Bound on ξ_Ω We have the following bound on ξ_Ω for "degree-bounded" Ω.

Lemma 3. Let $\text{supp}(A^*) \subseteq \{1,\ldots,n\}\times\{1,\ldots,n\}$ *be some fixed support set in which every row/column has at most δ non-zero entries. Then for Ω as defined by (3) and ξ_Ω defined by (9), we have that $\xi_\Omega \leq \delta$.*

Therefore, matrices that are "degree-bounded" have small spectral norm (compared to n).

Bound on ξ_T Before analyzing ξ_T we define the following notion of incoherence, which played an important role in the results of Candes and Recht [2008]:

$$\mu = \max(\max_i \|P_U e_i\|_2, \max_i \|P_V e_i\|_2), \tag{11}$$

where $B^* = U\Sigma V'$ and each $e_i \in \mathbb{R}^n$ is a standard basis vector. We provide a bound on ξ_T in terms of μ.

Lemma 4. Let T be the linear space defined in (4), let ξ_T be defined by (8), and let μ be defined by (11). Then,

$$\xi_T \leq 3\mu.$$

Conditions based on δ, μ Combining Propositions 1 and 2, and Lemmas 3 and 4, we have the following corollary.

Corollary 5. Suppose that $C = A^* + B^*$, and suppose that each row/column of A^* has at most δ non-zero entries. Then there exists a range of values for γ (given by (10)) so that (A^*, B^*) is the unique *optimizer of (2) if*

$$\delta\mu < \frac{1}{24},$$

where μ is defined by (11).

5.2 Exact recovery for random matrices

Next, we describe classes of random matrices such that if A^* and B^* are drawn from these classes, then we have exact recovery with very high probability. In particular, we show that the conditions specified by Corollary 5 are satisfied with high probability by matrices randomly drawn from these classes.

Random orthogonal model (Candes and Recht [2008]) The matrix $B^* = U\Sigma V'$ is constructed as follows: The singular vectors $U, V \in \mathbb{R}^{n\times k}$ are drawn *uniformly* at random from the collection of rank-k partial isometries in $\mathbb{R}^{n\times k}$. The singular vectors in U and V need not be independent from each other. No restriction is placed on the singular values.

Random sparsity model The matrix A^* is such that $\text{supp}(A^*)$ is chosen uniformly at random from the collection of all support sets of size m. There is no assumption made about the values of A^* at locations specified by $\text{supp}(A^*)$.

For the random orthogonal model, the value of μ was computed by Candes and Recht [2008].

Lemma 6. Suppose that a rank-k B^* is drawn according to the random orthogonal model. Then there exists a constant α_1 such that μ (defined by (11)) is bounded as

$$\mu \leq \alpha_1 \sqrt{\frac{\max(k, \log(n))}{n}},$$

with very high probability.

Finally, we obtain a degree bound on matrices A^* drawn according to the random sparsity model.

Lemma 7. Suppose that A^* is drawn according to the random sparsity model with m non-zero entries. Let δ be the maximum number of non-zero entries in each row/column of A^*. We have that

$$\delta \leq \frac{m}{n} \log(n),$$

with high probability.

The proof of this lemma follows from a standard balls and bins argument, and can be found in several references (see for example (Bollobas [2001])). Combining the previous two lemmas with Corollary 5, we have the following result for exact recovery for matrices drawn from the random orthogonal model and the random sparsity model.

Theorem 8. Suppose that a rank-k B^* is drawn from the random orthogonal model, and that A^* is drawn from the random sparsity model with m non-zero entries. Given $C = A^* + B^*$, there exists a range of values for γ (given by (10)) so that we can recover (A^*, B^*) exactly with high probability using the SDP (2) provided

$$m \leq \alpha \frac{n^{1.5}}{\log(n)\sqrt{\max(k, \log(n))}},$$

for some constant α.

Thus, for matrices B^* with rank k smaller than n, the SDP (2) yields exact recovery with high probability even when the size of the support of A^* is super-linear in n.

6. SIMULATION RESULTS

We describe two sets of experiments to evaluate the performance of our convex program (2) in recovering sparse and low-rank matrices. All simulations were performed using YALMIP [2004] and the SDPT3 software for solving SDPs (Toh et al. [2006]).

In the first experiment, we generate symmetric positive-definite rank-k matrices B^* as follows. We construct a matrix $P \in \mathbb{R}^{n \times k}$ with i.i.d Gaussian entries, and let $B^* = PP'$. To generate sparse symmetric matrices A^*, we choose a symmetric support set $\text{supp}(A^*)$ of size m uniformly at random. The values of A^* within this support are i.i.d Gaussian. Letting $C = A^* + B^*$, we solve the SDP (2) and conclude that recovery is successful if the solution $(\hat{A}, \hat{B})$ satisfies $\text{tol}_\gamma \leq 10^{-3}$ with

$$\text{tol}_\gamma = \frac{\|\hat{A} - A^*\|_F}{\|A^*\|_F} + \frac{\|\hat{B} - B^*\|_F}{\|B^*\|_F} \qquad (12)$$

for some[3] γ. The following table shows the success rate of exact recovery with $n = 20$ for various values of m, k with the above procedure repeated 50 times:

	m=30	m=35	m=40	m=45
k=3	98%	96%	90%	82%
k=6	54%	46%	22%	6%

Thus, for reasonably low-rank matrices B^* and for sufficiently sparse A^*, (2) recovers these matrices from $C = A^* + B^*$ exactly.

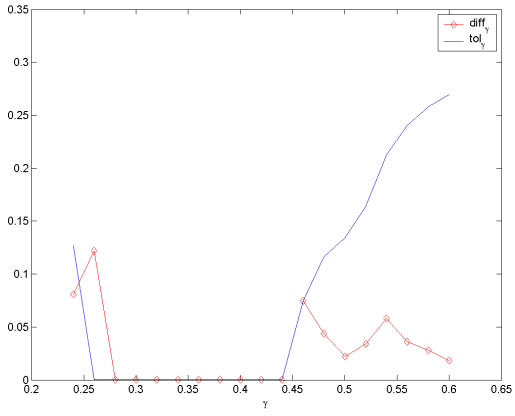

Fig. 2. Comparison between diff_γ and tol_γ for a randomly generated model with $n = 30, k = 6, m = 100$.

Next, we discuss an experiment that suggests a heuristic for choosing the regularization parameter γ in (2). We generate a single random instance of B^* and A^* by the same procedure as above, but with $n = 30, k = 6, m = 100$. We solve the convex program (2) for numerous values of γ, beginning with $\gamma = 0.25$ and incrementing the value of γ by $\epsilon = 0.02$. For each γ, we compute

$$\text{diff}_\gamma = \frac{\|\hat{A}_{\gamma-\epsilon} - \hat{A}_\gamma\|_F}{\|\hat{A}_\gamma\|_F} + \frac{\|\hat{B}_{\gamma-\epsilon} - \hat{B}_\gamma\|_F}{\|\hat{B}_\gamma\|_F}. \qquad (13)$$

We also compute the quantity tol_γ of (12) in order to check if A^*, B^* are recovered successfully. Figure 2 shows the results of this experiment. Notice that whenever the experiment is successful (i.e., when tol_γ is small), the value of diff_γ also tends to be zero. If the experiment is not successful, diff_γ typically does not tend to be close to zero and in fact changes rapidly. Therefore, if a good guess for γ is not available, one could solve (2) for a range of γ's and choose a solution corresponding to a γ in a range (γ_1, γ_2) in which diff_γ is stable and near zero.

7. DISCUSSION

We presented a method for decomposing a matrix $C = A^* + B^*$ into its components A^*, B^*, where A^* is sparse and B^* is low-rank. Such a problem arises in a variety of situations in statistical model selection and in system identification. Our approach is based on a convex optimization formulation that can be solved efficiently using SDP solvers. We provide a set of sufficient conditions that guarantee exact recovery of A^*, B^*. These conditions essentially require that the row-space and column-space of the low-rank matrix B^* are "incoherent" with respect to the standard basis, and that A^* does not have too many

[3] We perform our experiment for a large range of values of γ and conclude that recovery is exact if (12) is satisfied for some γ in this range. We typically found that whenever exact recovery was possible, values of γ around 0.4 provided the right level of trade-off.

non-zeros in each row/column. Our sufficient conditions can also be used to conclude that when A^* and B^* are drawn from certain natural random ensembles, the SDP (2) succeeds in exact recovery with high probability.

Our analysis extends beyond the results presented here. In the model selection problem presented in Section 2 the graphical connections between latent and observed variables may be sparse (i.e., the matrix $J_{o,h}$ in (1) may be sparse). This may potentially pose problems in terms of the incoherence conditions described in Section 5 being satisfied, as we require that the row and column spaces of B^* (in this case $B^* = -J_{o,h}J_h^{-1}J_{h,o}$ from (1)) must be incoherent with the coordinate axes. However, one can overcome this problem using the graph-theoretic concept of an *expander* (Hoory et al. [2006]) to describe connections between latent and observed variables in order to impose a minimal level of incoherence. In addition, our analysis can also be used to develop an "uncertainty principle" that characterizes rank-sparsity incoherence for arbitrary matrices using the quantities ξ_Ω and ξ_T defined by (3) and (4). We will discuss more details in a longer report.

An important problem for further research is that of approximately decomposing a matrix into sparse and low-rank components when the specified matrix C is not exactly of this form (for example, due to noise). Another interesting question is that of exploiting structure in the SDP (2) in order to provide for a more efficient solution than that provided by generic solvers, which may not exploit special structure.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Benjamin Recht for helpful discussions.

REFERENCES

D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientic, Belmont, MA, 2003.

B. Bollobas. *Random graphs*. Cambridge University Press, 2001.

E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, volume 52, number 2, pages 489–509, 2006.

E. J. Candes and B. Recht. Exact Matrix Completion Via Convex Optimization. *Submitted for publication*, 2008.

D.L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, volume 59, issue 6, pages 797–829, 2006.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, volume 52, number 4, pages 1289–1306, 2006.

D. L. Donoho and M. Elad. Optimal Sparse Representation in General (Nonorthogonal) Dictionaries via ℓ_1 Minimization. *Proceedings of the National Academy of Sciences*, volume 100, pages 2197–2202, 2003.

M. Fazel. Matrix Rank Minimization with Applications. *PhD thesis*, Department of Electrical Engineering, Stanford University, 2002.

S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, volume 43, number 4, pages 439–561, 2006.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

M. Mesbahi and G. P. Papavassilopoulos. On the rank minimization problem over a positive semidefinite linear matrix inequality. *IEEE Transactions on Automatic Control*, volume 42, number 2, pages 239-243, 1997.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed Minimum Rank Solutions to Linear Matrix Equations via Nuclear Norm Minimization. *Submitted to SIAM Review*, 2007.

K. C. Toh, M.J. Todd, and R. H. Tutuncu. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming*. Available from http://www.math.nus.edu.sg/ mattohkc/sdpt3.html.

L. G. Valiant. Graph-theoretic arguments in low-level complexity. *6th Symposium on Mathematical Foundations of Computer Science*, pages 162-176, 1977.

L. Vandenberghe and S. Boyd. Semidefinite Programming. *SIAM Review*, volume 38, number 1, pages 49–95, 1996.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*, volume 170, pages 1039–1053, 1992.

YALMIP: A Toolbox for Modeling and Optimization in MATLAB. *Proceedings of the CACSD Conference*, Taiwan, 2004. URL http://control.ee.ethz.ch/ joloef/yalmip.php.

Appendix A. SEMIDEFINITE FORMULATION

The problem (2) can be recast as a *semidefinite program* (SDP). We appeal to the fact that the spectral norm $\|\cdot\|$ is the dual norm of the nuclear norm $\|\cdot\|_*$:

$$\|M\|_* = \max\{\mathrm{trace}(M'Y) |\ \|Y\| \leq 1\}.$$

Further, the spectral norm admits a simple semidefinite characterization (Recht et al. [2007]):

$$\|Y\| = \min_t \quad t \qquad s.t. \begin{pmatrix} tI_n & Y \\ Y' & tI_n \end{pmatrix} \succeq 0.$$

The dual of this SDP is

$$\|M\| = \min_{W_1,W_2} \quad \frac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2))$$
$$s.t. \qquad \begin{pmatrix} W_1 & M \\ M' & W_2 \end{pmatrix} \succeq 0.$$

Putting these facts together, (2) can be rewritten as

$$\min_{A,B,W_1,W_2,Z} \quad \gamma\mathrm{trace}(Z\mathbf{1}_{n\times n}) + \frac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2))$$
$$s.t. \qquad \begin{pmatrix} W_1 & B \\ B' & W_2 \end{pmatrix} \succeq 0$$
$$-Z_{i,j} \leq A_{i,j} \leq Z_{i,j},\ \forall (i,j)$$
$$A + B = C. \qquad \text{(A.1)}$$

Here, $\mathbf{1}_{n\times n} \in \mathbb{R}^{n\times n}$ refers to the matrix that has 1 in every entry.