Noisy estimation of simultaneously structured models: Limitations of convex relaxation

Samet Oymak[c], Amin Jalali[w], Maryam Fazel[w], Babak Hassibi[c]

[c] California Institute of Technology
[w] University of Washington

***Abstract*— Models or signals exhibiting low dimensional behavior (e.g., sparse signals, low rank matrices) play an important role in signal processing and system identification. In this paper, we focus on models that have multiple structures simultaneously; e.g., matrices that are both low rank *and* sparse, arising in phase retrieval, quadratic compressed sensing, and cluster detection in social networks. We consider the estimation of such models from observations corrupted by additive Gaussian noise. We provide tight upper and lower bounds on the mean squared error (MSE) of a convex denoising program that uses a combination of regularizers to induce multiple structures. In the case of low rank and sparse matrices, we quantify the gap between the MSE of the convex program and the best achievable error, and we present a simple (nonconvex) thresholding algorithm that outperforms its convex counterpart and achieves almost optimal MSE.**

This paper extends prior work on a different but related problem: recovering simultaneously structured models from noiseless compressed measurements, where bounds on the number of required measurements were given. The present work shows a similar fundamental limitation exists in a statistical denoising setting.

***Index Terms*— simultaneously structured, low rank and sparse, denoising, estimation, compressed sensing**

I. Introduction

Suppose we would like to identify an unknown system or model, given limited and/or noisy observations of the system, and with the prior information that the system has a low-dimensional structure. To do this, one can setup an optimization problem that minimizes the error for fitting the observations, and penalizes (an appropriate notion of) model complexity. This setup is nowadays commonplace in signal processing, statistical learning, and dynamical system identification. Recent research has focused on theoretical analysis of classes of *structures* and their corresponding convex penalties, e.g., sparsity of coefficients induced by the ℓ_1 norm, or low-rankness of a matrix induced by the nuclear norm [3], [4], [14], [10]. As an example in linear dynamical systems identification, system order or McMillan degree is captured by the rank of an appropriate Hankel matrix, and penalties encouraging a low-rank solution have been employed [26].

Email: {soymak,bhassibi}@caltech.edu, {amjalali,mfazel}@uw.edu. Research supported in part by the National Science Foundation Career award ECCS-0847077 and by the NSF grants CCF-0729203, CNS-0932428 and CCF-1018927.

We consider the problem of *estimation* for structured models [1], [2], [13]: how well can a signal $\mathbf{x}_0$ be estimated from corrupted observations $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$, where $\mathbf{z}$ is additive noise? The aim is to find a signal that both fits the observations and satisfies the structural requirements, such as sparsity. Various estimation problems involving sparse vectors, low rank matrices and sum of low rank and sparse matrices have been analyzed in [1], [6], [11], [12]. Existing work has focused on models with a *single structure* only, whereas in practice, a model often has more than one structure at the same time. For example, simultaneously sparse *and* low-rank matrices arise in applications such as sparse phase retrieval (see, e.g., [21], [25], [24]), as well as cluster detection [7], [8]. The paper [9] proposed a general framework for recovery of such signals from linear observations using a combination of known penalties corresponding to each structure. It showed that, surprisingly, no combination of the individual penalties can do any better, order-wise, than an algorithm which uses only the *best* individual penalty and which essentially exploits only one of the several structures. The papers [20], [30] are in similar nature to [9].

The present paper focuses on the estimation problem for simultaneously structured signals and considers the convex program,

$$\arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \sum_{i=1}^{S} \lambda_i \|\mathbf{x}\|_{(i)} \tag{I.1}$$

to estimate $\mathbf{x}_0$ where the penalty function $\|\cdot\|_{(i)}$ is chosen to exploit the structure i of $\mathbf{x}_0$. We give new sharp bounds on the performance of (I.1) where we use mean-squared-error (MSE) as the performance metric. Specifically,

- In Section III, we analyze the problem for *abstract* signals. In a similar flavor to [9], [20], [30], our lower bounds on the MSE indicate that using a combination of penalties may not do a much better job at estimation than using only the best penalty.
- In Section IV, we specialize our results to the case of *low rank and sparse matrices* where the structures are sparsity, row/column sparsity and low rankness. While this has been studied in [7], authors only provide upper bounds to the estimation error. Our analysis are more comprehensive and we additionally provide a sharp MSE lower bound which *provably* verifies the belief that one doesn't gain much by using multiple penalties

in this special case. In particular, we show this in the high signal-to-noise ratio (SNR) regime.
- In Section V, for sparse and low rank matrices, we propose a simple and tractable algorithm, based on thresholding the nonzero entries and singular values of the matrix. We show that this approach can substantially *outperform* the convex program in the same SNR regime.

Section VI is dedicated to verify our theoretical predictions and contrasting convex and nonconvex algorithms. Finally, Section VII discusses open questions and avenues for future research.

II. Problem Setup

Notation. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|$ denotes a general norm, and $\|\mathbf{x}\|^* = \sup_{\|z\|\leq 1} \langle \mathbf{x}, \mathbf{z} \rangle$ denotes the dual norm. A subgradient of the norm $\|\cdot\|$ at $\mathbf{x}$ is a vector $\mathbf{g}$ for which $\|\mathbf{z}\| \geq \|\mathbf{x}\| + \langle \mathbf{g}, \mathbf{z}-\mathbf{x} \rangle$ holds for any $\mathbf{z}$. The set of all subgradients is called the subdifferential, denoted by $\partial\|\mathbf{x}\|$. For a subspace M, denote its orthogonal complement by $M^\perp$. We use $\mathbf{x}_M$ as a shorthand for the orthogonal projection of $\mathbf{x}$ onto subspace M defined as

$$\mathbf{x}_M = \arg\min_{\mathbf{w}\in M} \|\mathbf{w}-\mathbf{x}\|_2 .$$

The ℓ_1 norm of a vector or matrix is denoted by $\|\cdot\|_1$ and returns the sum of the absolute values of its entries. The $\ell_{1,2}$ norm of a matrix is denoted by $\|\cdot\|_{1,2}$ and returns the sum of the ℓ_2 norms of the columns, i.e., $\|\mathbf{X}\|_{1,2} = \sum_i \|\mathbf{x}_i\|_2$ where $\mathbf{x}_i$ is the i'th column, [11]. Also, $\|\cdot\|_{0,2}$ denotes the nonconvex function that returns the number of nonzero columns of a matrix. Finally, the nuclear norm of a matrix $\mathbf{X}$ is denoted by $\|\mathbf{X}\|_\star$ and is equal to the sum of its singular values.

Remark: Due to space considerations, some of the proofs are omitted. The reader is referred to the technical report [31].

A. Simultaneous structures

We consider a signal (or model) $\mathbf{x}_0 \in \mathbb{R}^n$ that has several "low dimensional structures" simultaneously (e.g., sparsity, group sparsity, or low-rank). Suppose the i'th structure corresponds to a norm $\|\cdot\|_{(i)}$, that when used as a penalty, promotes that structure (e.g., ℓ_1 norm for sparsity). We refer to such a $\mathbf{x}_0$ a *simultaneously structured* model. As a central example, consider a simultaneously sparse and low rank matrix $\mathbf{X}_0 \in \mathbb{R}^{n_1\times n_2}$, where nonzero entries lie on a $k\times k$ submatrix, for some $k \ll n$, and rank is $r \ll k$. To induce sparsity and low rank, we can use the ℓ_1 norm and the nuclear norm, respectively, as penalties.

To simplify the analysis, in this paper we assume that all the norms are *decomposable at* $\mathbf{x}_0$ which is defined as follows.

Definition 2.1 (Decomposable Norm): A norm $\|\cdot\|$ is decomposable at $\mathbf{x} \in \mathbb{R}^n$ if there exist a subspace $T \subset \mathbb{R}^n$ and a vector $\mathbf{e} \in T$ such that the subdifferential at $\mathbf{x}$ has the form

$$\partial\|\mathbf{x}\| = \{\mathbf{z} \in \mathbb{R}^n \,:\, \mathbf{z}_T = \mathbf{e}\,,\ \|\mathbf{z}_{T^\perp}\|^* \leq 1\}, \tag{II.1}$$

and for all $\mathbf{s} \in T^\perp$ we have

$$\|\mathbf{s}\| = \sup_{\mathbf{z}\in T^\perp, \|\mathbf{z}\|^*\leq 1} \langle \mathbf{s}, \mathbf{z} \rangle . \tag{II.2}$$

We refer to T as the *support* and $\mathbf{e}$ as the *sign vector* of $\mathbf{x}_0$ with respect to $\|\cdot\|$.

The prime example is the ℓ_1 norm. It is easy to check that this norm is decomposable at vectors $\mathbf{x}_0$ with k nonzero entries (called k-sparse vectors). The sign vector $\mathbf{e}$ is given by entrywise signs, i.e., $\mathbf{e}_i = \operatorname{sgn}(\mathbf{x}_{0,i})$, and the support T is the subspace of vectors whose nonzero patterns are same as $\mathbf{x}_0$. The ℓ_1, $\ell_{1,2}$, and nuclear norms are decomposable at all points in $\mathbb{R}^n$.

B. The Denoising Problem

We consider the problem of estimating a simultaneously structured model (signal) $\mathbf{x}_0 \in \mathbb{R}^n$ corrupted by additive Gaussian noise $\mathbf{z}$, where we observe the corrupted signal $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$. Suppose $\mathbf{x}_0$ has S different known structures (e.g., sparsity, block sparsity, or low rank), and that the i'th structure is associated with the decomposable norm $\|\cdot\|_{(i)}$. To estimate $\mathbf{x}_0$, we consider the following optimization problem,

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \sum_{i=1}^{S} \lambda_i \|\mathbf{x}\|_{(i)}, \tag{II.3}$$

where $\hat{\mathbf{x}}(\mathbf{y})$ stands for the estimate obtained via (II.3) when the observed vector is $\mathbf{y}$, and λ_i are regularization parameters. As a performance measure, we use the usual mean squared error (MSE), $\|\hat{\mathbf{x}}-\mathbf{x}_0\|_2^2$; to obtain a measure independent of noise statistics, we normalize the MSE by the noise variance. The following definition characterizes the worst-case estimation performance of the program (II.3) at $\mathbf{x}_0$.

Definition 2.2 (MSE risk): Consider the denoising program (II.3) with regularization parameters $\lambda_1, \ldots, \lambda_S$ and noise vector $\mathbf{z}$ with i.i.d. standard normal entries. We define the MSE risk of the program (II.3) at $\mathbf{x}_0$ as

$$\eta(\{\lambda_i\}_{i=1}^S) = \max_{\gamma>0} \mathbb{E}[\|\hat{\mathbf{x}}(\gamma\mathbf{x}_0+\mathbf{z}) - \gamma\mathbf{x}_0\|_2^2]. \tag{II.4}$$

That is, we consider scaled versions of $\mathbf{x}_0$ while keeping the noise constant, and calculate the "worst case mean squared error" [15], [16], [17]. The MSE risk is a function of the signal $\mathbf{x}_0$, norms $\|\cdot\|_{(i)}$, and $\{\lambda_i\}_{i=1}^S$. Observe that, scaling does not change how structured the signal is; rank, sparsity or block sparsity of $\gamma\mathbf{x}_0$ is the same as $\mathbf{x}_0$ for nonzero γ. Thus $\eta(\{\lambda_i\}_{i=1}^S)$ provides a simple characterization of the worst case MSE independent of the scaling. Note that instead of scaling $\mathbf{x}_0$, we can obtain the same error measure by keeping $\mathbf{x}_0$ fixed, scaling the noise vector and $\{\lambda_i\}_{i=1}^S$ by σ, then normalizing the error by σ^2. This is because $\|\cdot\|_{(i)}$ are homogeneous functions. Thus, we can think of γ as the signal-to-noise ratio (SNR). We will see in Section III that this error measure captures the fundamental properties of structured signal estimation.

• **Known results:** Sparse signals and low rank matrices are two important cases where the desired model has a single structure. When $\mathbf{x}_0 \in \mathbb{R}^n$ is k-sparse, with the right tuning of ℓ_1 norm, the MSE risk is as small as $\mathcal{O}\left(k\log\frac{n}{k}\right)$. This

differs from the minimum possible MSE risk, which is k, by only a log-factor [1]. Here, by "minimum possible", we mean the smallest MSE risk over all estimators which includes nonconvex functions such as cardinality ℓ_0. The denoiser (estimator) with ℓ_1 norm is also known as the soft thresholding operator. For a rank r matrix in $\mathbb{R}^{n\times n}$, the MSE risk of properly tuned nuclear norm minimization is $\mathcal{O}(rn)$ and orderwise optimal, [13], [18], [19].

• **In this work**, we provide bounds for the MSE risk of (II.3) for simultaneously structured signals. In the case of sparse and low rank matrices, our lower bound indicates a fundamental limitation of using a combination of ℓ_1 norm and the nuclear norm for estimation: As an example, assume $\mathbf{X}_0$ is a rank 1 matrix whose entries are contained in a $k\times k$ submatrix. While a mean squared error of $\mathcal{O}(k)$ is achievable via a simple nonconvex algorithm, we show that, the risk for problem (II.3) is lower bounded by $\mathcal{O}\left(\min\{k^2,n\}\right)$, which can be significantly higher than k.

III. Bounds on MSE

In this section, we provide several results on the performance of program (II.3), including upper and lower bounds on the mean squared error and a formula that yields the exact MSE risk (2.2). We first present a slight modification of a result from [15] that gives the MSE risk in terms of the sub differentials of the norms at $\mathbf{x}_0$.

Theorem 3.1 (Exact MSE risk, [15]): Consider the program (II.3) with noise vector $\mathbf{z}\sim\mathcal{N}(0,I)$ and regularization parameters $\{\lambda_i\}_{i=1}^S$.

- Recall Definition 2.2. The MSE risk is given as follows,

$$\eta(\{\lambda_i\}_{i=1}^S)=\mathbb{E}[\text{dist}(\mathbf{z},\sum_i\lambda_i\partial\|\mathbf{x}_0\|_{(i)})^2], \tag{III.1}$$

  where $\text{dist}(\mathbf{z},C)$ denotes the Euclidean distance of a point $\mathbf{z}$ to the set C.
- Furthermore, the equality in (II.4) is achieved as $\gamma\to\infty$, i.e. the scaling that maximize the MSE is infinity.

Since MSE risk is the worst case mean squared error over all scalings, the equation (III.1) also provides an upper bound to the mean squared error of program II.3.

$$\mathbb{E}[\|\hat{\mathbf{x}}(\mathbf{x}_0+\mathbf{z})-\mathbf{x}_0\|_2^2]\le\mathbb{E}[\text{dist}(\mathbf{z},\sum_i\lambda_i\partial\|\mathbf{x}_0\|_{(i)})^2] \tag{III.2}$$

Next, we will give a sharp lower bound for the MSE risk which is easier to interpret.

A. Lower bound on MSE

Recall the Definition 2.1. Let T_i and $\mathbf{e}_i$ be the support and sign vector corresponding to $\|\cdot\|_{(i)}$. Further, let $R_i=\sup_{\mathbf{s}\in\partial\|\mathbf{x}_0\|_{(i)}}\|\mathbf{s}\|_2$ which is the largest subgradient of the norm at $\mathbf{x}_0$. Finally, let $T_\cap=\bigcap_{i=1}^S T_i$ denote the *joint support*. We define a quantity that will be useful for the next result. Let

$$\zeta_i=\frac{\|(\mathbf{e}_i)_{T_\cap}\|_2^2\,n}{R_i^2\dim(T_i)}, \tag{III.3}$$

where $(\mathbf{e}_i)_{T_\cap}$ is the orthogonal projection of vector e_i onto the subspace $T_\cap$. We can view $\{\zeta_i\}_{i=1}^S$ as a measure of correlation among the subgradients of the norms over the joint support subspace $T_\cap$. As argued in Section 2 of [9], often (and for all norms mentioned in this paper) ζ_i lies between 0 and 1.

Theorem 3.2 (General lower bound): Consider the MSE risk (III.1). Let ζ_i be as in (III.3) and assume

$$\left\langle(\mathbf{e}_i)_{T_\cap},(\mathbf{e}_j)_{T_\cap}\right\rangle\ge 0, \tag{III.4}$$

for all $i,j\le S$. Finally, assume the model is reasonably low-dimensional in at least one of the structures, namely, that $\min_i\zeta_i\dim(T_i)\le n$. Then, the MSE risk given in (III.1) satisfies

$$\eta(\{\lambda_i\}_{i=1}^S)\ge\frac{1}{6S}\min_{1\le i\le S}\zeta_i\dim(T_i). \tag{III.5}$$

The proof of this result can be found in [31] and it is in a similar nature to the proof of Theorem 3.2 of [9].

Interpretation: To have an intuitive interpretation, we can view the dimension of support subspace T_i as the complexity or "degrees of freedom" of the signal under structure i. For example, for a k sparse signal, the support corresponding to ℓ_1 norm is k dimensional. Ignoring the ζ_i terms (since as mentioned before for all norms we consider they are bounded by 1), Theorem 3.2 suggests the MSE risk is at least on the same order as the dimension of the smallest support subspace. Incorporating more norms for estimation, may reduce the MSE risk, but only by a constant factor $\frac{1}{S}$. It is worth noting that (III.5) is independent of $\{\lambda_i\}_{i=1}^S$.

Lower bounding MSE: Recall that MSE risk is the worst case MSE. This means, for any $\varepsilon>0$, for sufficiently large SNR γ, the MSE in estimating $\gamma\mathbf{x}_0$ from $\gamma\mathbf{x}_0+\mathbf{z}$ will be greater than $\frac{1-\varepsilon}{6S}\min_{1\le i\le S}\zeta_i\dim(T_i)$.

B. Upper bound on MSE

While Section III-A finds a lower bound on the MSE risk, next, we give a *deterministic* upper bound to the estimation error in the general decomposable norm framework.

Theorem 3.3: Let $\mathbf{x}_0$ be a simultaneously structured model corresponding to $\{\|\cdot\|_{(i)}\}_{i=1}^S$. Define the constants,

$$\rho_i\triangleq\sup_{\mathbf{v}}\{\|\mathbf{v}_{T_i}\|_2:\ \|\mathbf{v}\|_{(i)}^*\le 1\}. \tag{III.6}$$

for $1\le i\le S$ corresponding to the i'th norm at $\mathbf{x}_0$ (see Definition 2.1). Let $\mathbf{z}$ be the noise vector and $\hat{\mathbf{x}}$ be the estimate obtained from program (II.3) for $\mathbf{y}=\mathbf{x}_0+\mathbf{z}$. Assume for some set of non-negative constants $\{\bar{\lambda}_i\}_{i=1}^S$ with $\sum_i\bar{\lambda}_i=1$, we have $\lambda_i\ge\bar{\lambda}_i\|\mathbf{z}\|_{(i)}^\star$. Then, the error satisfies,

$$\|\hat{\mathbf{x}}-\mathbf{x}_0\|_2^2\le\min\left\{4\sum_{i=1}^S\lambda_i\|\mathbf{x}_0\|_{(i)}\ ,\ \left(4\sum_{i=1}^S\lambda_i\rho_i\right)^2\right\}.$$

ρ_i is closely related to $\dim(T_i)$ and can be seen as a measure of complexity of the signal under norm $\|\cdot\|_{(i)}$. Theorem 3.3 generalizes Proposition 1 of [7] to arbitrary mixtures of decomposable norms.

In the case of a simultaneously sparse and low rank matrix $\mathbf{X}_0$ with $k\times k$ nonzero elements and rank r, we have the following.

- For sparsity: Use ℓ_1 norm, $\rho_{\ell_1}=k$.

- For low rank: Use nuclear norm, $\rho_\star = \sqrt{2r}$.

Using these, we can obtain upper bounds to the MSE which are in similar nature to the ones given by Richard et al [7].

IV. MSE of Simultaneously Sparse and Low Rank Matrices

We now specialize our general bounds to the simultaneously sparse and low rank (S&L) matrices. Let us first start with a description of our model.

Definition 4.1 (S&L): Let $\mathbf{X}_0 \in \mathbb{R}^{n\times n}$ be a rank r matrix. Assume the smallest submatrix containing nonzero entries of $\mathbf{X}_0$ has dimensions $k\times k$.

Observe that when $k \ll n$ the matrix becomes sparse and if additionally $r \ll k$, the matrix is low rank with respect to the nonzero $k\times k$ submatrix. Hence, it exhibits two structures. Observe that S&L matrices additionally have column and row sparsity. Hence, the ℓ_1 norm, $\ell_{1,2}$ norm and the nuclear norm are strong candidates for the structure inducing functions.

A. Lower bound on S&L MSE

Theorem 4.1 (S&L lower bound): Let $\mathbf{X}_0 \in \mathbb{R}^{n\times n}$. Consider the following cases and the associated MSE risk.

(a) Assume $\mathbf{X}_0$ is a rank 1 matrix satisfying $\mathbf{X}_0 = \mathbf{a}\mathbf{b}^T$ where $\mathbf{a},\mathbf{b}$ are unit norm k sparse vectors. Then, MSE risk of the program

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 + \lambda_{\ell_1}\|\mathbf{X}\|_1 + \lambda_\star\|\mathbf{X}\|_\star \tag{IV.1}$$

at $\mathbf{X}_0$ obeys $\eta(\lambda_{\ell_1},\lambda_\star) \geq \frac{1}{12}\min\{\|\mathbf{a}\|_1\|\mathbf{b}\|_1 k,\, n\}$.

(b) Assume $\mathbf{X}_0$ is same as in Definition 4.1. The MSE risk of the program

$$\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 + \lambda_{\ell_{1,2}}\|\mathbf{X}\|_{1,2} + \lambda_{\ell_{1,2}^T}\|\mathbf{X}^T\|_{1,2} + \lambda_\star\|\mathbf{X}\|_\star \tag{IV.2}$$

at $\mathbf{X}_0$ is lower bounded by $\frac{1}{18}rn$.

Proof: Observe that, to apply Theorem 3.2, one needs the inner products of the projected sign vectors to be nonnegative. For S&L matrices, this was shown in [9], Section 5. For (IV.1), results in [9] give $\zeta_{\ell_1} = \frac{\|\mathbf{a}\|_1\|\mathbf{b}\|_1}{k}, \zeta_\star = 1$ and $\dim(T_{\ell_1}) = k^2, \dim(T_\star) = 2n-1 \geq n$ which yields the desired result with $S=2$. For (IV.2), results in [9] give $\zeta_{\ell_{1,2}} = \zeta_\star = 1$ and $\dim(T_{\ell_{1,2}}) = kn \geq rn, \dim(T_\star) = r(2n-r) \geq rn$ which again gives the desired outcome by setting $S=3$. ∎

Remark: In (a), when entries of $\mathbf{a}$ and $\mathbf{b}$ are $\pm 1/\sqrt{k}$, the lower bound takes a simpler form, namely $\eta \geq \frac{\min\{k^2,n\}}{12}$. What is remarkable is that this result shows using the best possible combination is no better than using only one of the ℓ_1 or nuclear norms.

B. Upper bound on S&L MSE

Lemma 4.2: Let noise $\mathbf{Z}$ have i.i.d. standard normal entries. Let $\mathbf{X}_0$ be same as in Definition 4.1. Then, for the optimal choice of λ_{ℓ_1} and $\lambda_\star$ (i.e. the choice that minimize the MSE risk), the MSE risk of the program (IV.1) at $\mathbf{X}_0$ satisfies

$$\eta(\lambda_{\ell_1},\lambda_\star) \leq \min\{c_1 k^2 \log\frac{n}{k},\, c_2 n\}. \tag{IV.3}$$

Similarly, for the optimal choice of $\lambda_{\ell_{1,2}}, \lambda_{\ell_{1,2}^T}, \lambda_\star$, the MSE risk of program (IV.2) at $\mathbf{X}_0$ satisfies,

$$\eta(\lambda_{\ell_{1,2}}, \lambda_{\ell_{1,2}^T}, \lambda_\star) \leq c_2 nr \tag{IV.4}$$

for some positive global constants c_1 and c_2.

Proof: To upper bound the minimum achievable MSE risk, we will consider the optimal choice for only one of the parameters and set the other one to zero. Consider the first statement. If $\lambda_{\ell_1} = 0$, $\lambda_\star$ can be tuned to obtain MSE of $c_2 rn = c_2 n$, [13]. Similarly, by setting $\lambda_\star = 0$ and optimally tuning λ_{ℓ_1}, one can obtain $c_1 k^2 \log\frac{n^2}{k^2}$ ([1]), which gives the desired result. For the second statement, the bound can be achieved by only tuning the nuclear norm and setting the block sparsity regularization parameters to 0. ∎

C. Combining the bounds

Combining Lemma 4.2 with Theorem 4.1, we obtain tight upper and lower bounds on the risk of optimally tuned estimation of low rank and sparse matrices. Namely, we have the following corollary.

Corollary 4.3: Assume $\mathbf{X}_0$ is as in Theorem 4.1 (a). Let $\lambda_{\ell_1}^*, \lambda_\star^*$ be chosen to minimize the MSE risk of (IV.1), i.e., $\eta^* = \eta(\lambda_{\ell_1}^*, \lambda_\star^*) = \min_{\lambda_{\ell_1}\geq 0, \lambda_\star \geq 0} \eta(\lambda_{\ell_1}, \lambda_\star)$. Then η^* satisfies

$$\frac{1}{12}\min\{\|\mathbf{a}\|_1\|\mathbf{b}\|_1 k, n\} \leq \eta^* \leq \min\{c_1 k^2 \log\frac{n}{k},\, c_2 n\}. \tag{IV.5}$$

Assume $\mathbf{X}_0$ is as in Theorem 4.1 (b) and define minimum MSE risk η^* similarly, with $\lambda_{\ell_{1,2}}, \lambda_{\ell_{1,2}^T}, \lambda_\star$. Then η^* for the program (IV.2) satisfies

$$\frac{1}{18} rn \leq \eta^* \leq c_2 nr\,. \tag{IV.6}$$

Corollary 4.3 shows that we can sandwich the optimally tuned risk between tight upper and lower bounds. For example, the ratio between the bounds in (IV.5) is only a $O(\log\frac{n}{k})$ factor when $\|\mathbf{a}\|_1, \|\mathbf{b}\|_1 \approx O(\sqrt{k})$.

V. A two-step thresholding algorithm

In this section, we present a simple and tractable (but nonconvex) algorithm for denoising an S&L matrix. Algorithm 1 first estimates the nonzero support of $\mathbf{X}_0$ and then uses singular value soft thresholding, [13]. Our goal is to contrast the performance of this algorithm with the convex programs (IV.1) and (IV.2). For Algorithm 1, we have the following result.

Theorem 5.1: Let $\mathbf{X}_0$ be same as in Definition 4.1 with rank $r=1$ and $\mathbf{Y} = \gamma\mathbf{X}_0 + \mathbf{Z}$ be the input to Algorithm 1. Assume $\mathbf{Z}$ has i.i.d. standard normal entries. There exists a scaling γ_0 and sparsity level k_0 such that for any $\gamma > \gamma_0$ and $k > k_0$, with probability $1 - 2\exp(-\mathscr{O}(k))$, we have,

$$\|\hat{\mathbf{X}}(\mathbf{Y}) - \gamma\mathbf{X}_0\|_F^2 \leq c_0 k \tag{V.2}$$

for some absolute constant $c_0 > 0$.

In words, Theorem 5.1 says that, if the signal to noise ratio γ is sufficiently large, Algorithm 1 will have an estimation error that grows linearly in k. This is a significant improvement over the convex program. In particular, combining

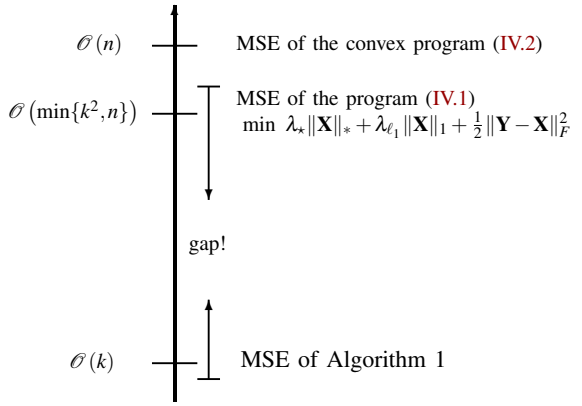
Algorithm 1 2-Step thresholding

Setup: $\mathbf{X}_0$ is same as in Definition 4.1.
Input: Noisy observation $\mathbf{Y} = \mathbf{X}_0 + \mathbf{Z} \in \mathbb{R}^{n\times n}$, sparsity k and rank r.
Output: $\hat{\mathbf{X}}(\mathbf{Y})$, estimate of $\mathbf{X}_0$.
Step 1: Find k rows and columns of $\mathbf{Y}$ with highest ℓ_2 norm. Set remaining columns and rows of $\mathbf{Y}$ to 0.
Step 2: Solve singular value soft thresholding,

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 + 2.01\sqrt{k}\|\mathbf{X}\|_\star \qquad \text{(V.1)}$$

Remark: In Step 2, the constant 2.01 is due to a technicality in our proof. In general, any constant above 2 will work.

second statement of Theorem 3.1 with the MSE risk obtained from Corollary 4.3 we find that, in the high SNR regime, the convex program's mean-squared-error is $\mathscr{O}\left(\min\{k^2,n\}\right)$ which can be significantly larger than k. The picture below illustrates this gap for a rank 1 matrix.



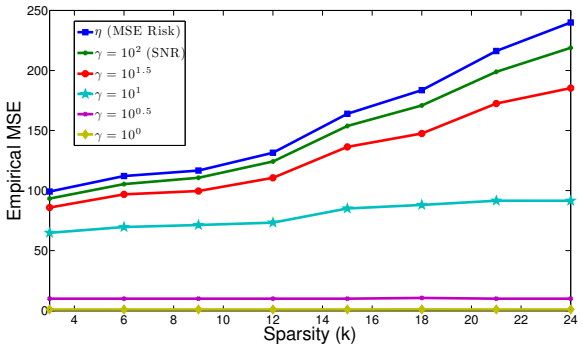
VI. Numerical results

Setup: The noise matrix $\mathbf{Z}$ has i.i.d. standard normal entries and we estimate $\gamma\mathbf{X}_0$ from $\gamma\mathbf{X}_0+\mathbf{Z}$ with the convex program (IV.1) or with Algorithm 1. We perform three experiments to confirm our theoretical predictions. The underlying matrix $\mathbf{X}_0$ is generated randomly as a rank one matrix $\frac{\mathbf{a}\mathbf{b}^T}{\|\mathbf{a}\mathbf{b}^T\|_F}$ where $\mathbf{a},\mathbf{b}\in\mathbb{R}^n$ are k sparse vectors whose nonzero entries are i.i.d. standard normal. For the plots of convex programs, results are averaged over 25 experiments. For Algorithm 1 this number is 200.

First, we aim to verify the risk formula given by Theorem 3.1. To do this, we solved (IV.1) for a 25×25 matrix, setting $(\lambda_{\ell_1},\lambda_\star) = (0.8,5)$ and increased the SNR γ from 10^0 to 10^2. The sparsity k is varied from 3 to 24. In Figure 1, as γ increases, the MSE increases as well and eventually becomes quite close to the exact MSE risk given by (III.1).

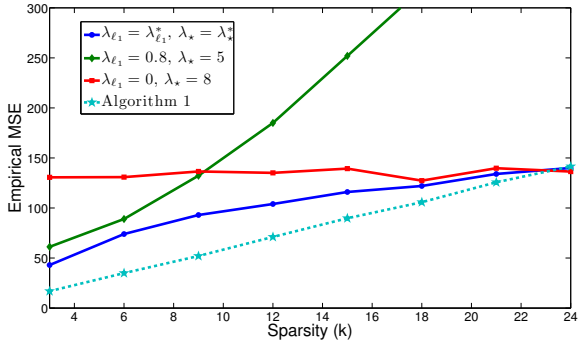
Next, we compared (IV.1) and Algorithm 1 for 30×30 matrices. We chose a high signal to noise ratio $\gamma^2 = 10^5$ and plotted the empirical mean squared errors corresponding to Algorithm 1 and (IV.1) with $(\lambda_{\ell_1},\lambda_\star) = (0,8),(0.8,5)$ and

Fig. 1. MSE of convex program for various SNRs ($\lambda_{\ell_1}=0.8,\ \lambda_\star=5$).



the optimal parameters $(\lambda^*_{\ell_1},\lambda^*_\star)$. $(\lambda^*_{\ell_1},\lambda^*_\star)$ are numerically estimated and they change as a function of k. In Figure 2, we observe that MSE of Algorithm 1 has a linear increase in k and it outperforms its convex counterparts in almost all sparsity regimes, and the difference is more significant for small k's.

Fig. 2. Algorithm 1 vs convex program (IV.1).



Finally, in Figure 3, we plotted the MSE of Algorithm 1 where $n=200$ and $k\in\{10,20,30\}$. We still observe that MSE is linear in k for large γ. When γ is small, both $\gamma\mathbf{X}_0$ and estimate of Algorithm 1's are small hence, resulting MSE is approximately zero. The interesting regime is when γ is between 1 and 10. In this regime, step 1 of the Algorithm fails to identify the correct nonzero rows that correspond to $\mathbf{X}_0$ which introduces additional error. This contrasts with the convex program, as in Figure 1, the MSE is a strictly decreasing function of the SNR γ.

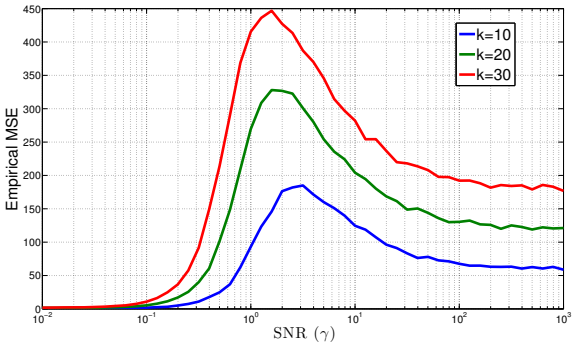
VII. Conclusions and Discussion

We studied the estimation of a low rank and sparse matrix corrupted by noise. Interestingly, we showed that estimation performance of an intuitive approach that combines individual convex penalties can be significantly worse than the best possible performance, while a simple nonconvex algorithm we propose achieves an MSE on the order of the best possible, in certain noise regimes.

A. Connection to Sparse PCA

The estimation of such a signal appears to be closely related to finding the sparse principal component of a matrix.

Fig. 3. Performance of Algorithm 1 as a function of SNR.



In sparse PCA, given the matrix $\mathbf{A}$, the aim is to find a unit vector $\mathbf{x}$ which maximizes $\mathbf{x}^T\mathbf{A}\mathbf{x}$ while ensuring $\mathbf{x}$ is sparse and has unit length, [27], [28], [29]. A convex relaxation for this problem is to lift $\mathbf{x}$ as $\mathbf{X} = \mathbf{x}\mathbf{x}^T$ (observe that $\mathbf{X}$ will be a low rank and sparse matrix), and solve the program

$$\begin{array}{ll} \max_{\mathbf{X}} & \langle \mathbf{X}, \mathbf{A} \rangle - \lambda \|\mathbf{X}\|_1 \\ \text{subject to} & \text{trace}(\mathbf{X}) = 1, \ \mathbf{X} \succeq 0. \end{array}$$

This program is proposed in [29], and similar to the denoising program IV.1, makes use of the ℓ_1 norm and the nuclear norm (or trace, for positive semidefinite matrices). We may view $\mathbf{A}$ as a matrix which is sum of a true low rank and sparse $\mathbf{x}_0\mathbf{x}_0^T$ and a noise term. Then, a term by term comparison between sparse PCA (VII-A) and the denoising program (IV.1) reveals that $\langle \mathbf{X}, \mathbf{A} \rangle$ corresponds to $\|\mathbf{X} - \mathbf{Y}\|_F^2$ and $\text{trace}(\mathbf{X}) = 1$ with $\mathbf{X} \succeq 0$ corresponds to fixing the nuclear norm $\|\mathbf{X}\|_\star$. Consequently, it would be interesting to analyze the performance of the sparse PCA program for the i.i.d. noise setup, and comparing with the results presented in this work.

B. Simultaneous LASSO

In many applications, observations of the signal $\mathbf{x}_0$ are given by a noisy linear map $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{z}$. A reasonable way to tackle this is to modify program II.3 as

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \sum_{i=1}^{S} \lambda_i \|\mathbf{x}\|_{(i)},$$

This problem, also known as the LASSO problem, can be seen as a combination of compression and denoising problems and has been analyzed for sparse vectors and low rank matrices extensively (see [2], [16], [10]). It is interesting to extend our bounds to the recovery error for this problem. This would generalize and merge the results of the present paper and [7], [9], [20], [30].

References

[1] DL Donoho, "Denoising by soft theresholding", IEEE Trans. Inform. Theory, 1995.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society, 58:267–288, 1996.

[3] E. J. Candès and T. Tao, "Decoding by linear programming," IEEE Trans. Inform. Theory, 51 4203-4215.

[4] D.L. Donoho, "Compressed sensing," IEEE Trans. Inform. Theory, 52(4):1289-1306, 2006.

[5] E. J. Candes, JK Romberg and T Tao, "Stable signal recovery from incomplete and inaccurate measurements". Comm. on Pure and Applied Math. Vol. 59, Issue 8, pg 12071223, August 2006.

[6] JF Cai, EJ Candès, and Zuowei Shen, "A Singular Value Thresholding Algorithm for Matrix Completion", SIAM J. on Optimization 20, 4 (March 2010), 1956-1982.

[7] E. Richard, PA Savalle, and N. Vayatis, "Estimation of Simultaneously Sparse and Low Rank Matrices", In Proceeding of 29th Annual International Conference on Machine Learning, 2012.

[8] BPW Ames, S. Vavasis. "Nuclear norm minimization for the planted clique and biclique problems", Math. Program. Ser. B (2011)129:69-89.

[9] S. Oymak, A. Jalali, M. Fazel, YC Eldar, and B. Hassibi, "Simultaneously Structured Models with Application to Sparse and Low-rank Matrices", arXiv:1212.3753.

[10] B. Recht, M. Fazel, P. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization". SIAM Review, Vol 52, no 3, pages 471-501, 2010.

[11] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, J. Roy. Stat. Soc. Ser. B Stat. Methodol., vol. 68, no. 1, pp. 4967, 2006.

[12] A. Agarwal, S. Negahban, M. J. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," Annals of Statistics, Volume 40, Number 2 (2012), 1171-1197.

[13] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear norm penalization and optimal rates for noisy matrix completion", Annals of Statistics, 2011.

[14] M. Fazel, "Matrix Rank Minimization with Applications". Elec. Eng. Dept, Stanford University, March 2002.

[15] S. Oymak and B. Hassibi, "Asymptotically Exact Denoising in Relation to Compressed Sensing", arXiv:1305.2714.

[16] M. Bayati and A. Montanari, "The LASSO Risk for Gaussian Matrices," IEEE Trans. on Info. Theory, VOL. 58, NO. 4, APRIL 2012.

[17] D. L. Donoho, I. Johnstone, and A. Montanari, "Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising", IEEE Trans. on Info. Theory, vol.59, no.6, pp.3396–3433, June 2013.

[18] D. L. Donoho and M. Gavish, "Minimax Risk of Matrix Denoising by Singular Value Thresholding", arXiv:1304.2085.

[19] Andrey Shabalin and Andrew Nobel, "Reconstruction of a Low-rank Matrix in the Presence of Gaussian Noise", arXiv:1007.4148.

[20] X. Li, V. Voroninski, "Sparse Signal Recovery from Quadratic Measurements via Convex Programming," arXiv:1209.4785.

[21] A. Walther. "The question of phase retrieval in optics". Opt. Acta, 10:41–49, 1963.

[22] R.P. Millane, "Phase retrieval in crystallography and optics". J. Opt. Soc. Am. A 7, 394-411 (1990).

[23] R.W. Harrison, "Phase problem in crystallography". J. Opt. Soc. Am. A, 10(5):1045–1055, 1993.

[24] E. J. Candès, Y.C. Eldar, T. Strohmer and V. Voroninski, "Phase retrieval via matrix completion", arXiv:1109.0573, Sep. 2011.

[25] R.W. Gerchberg and W.O. Saxton, "Phase retrieval by iterated projections", Optik 35, 237, 1972.

[26] M. Fazel, T. K. Pong, D. Sun, P. Tseng, "Hankel Matrix Rank Minimization with Applications to System Identification and Realization." SIAM Journal on Matrix Analysis and Applications, 34(3): 946-977, 2013.

[27] M. Journée, Y. Nesterov, P. Richtárik, and R Sepulchre, "Generalized Power Method for Sparse Principal Component Analysis". Journal of Machine Learning Research 11 (2010) 517–553.

[28] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis", Journal of computational and graphical statistics 15 (2), 265-286.

[29] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. "A Direct Formulation for Sparse PCA Using Semidefinite Programming". SIAM Rev. 49, 3 (July 2007), 434-448.

[30] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery", arXiv:1307.5870.

[31] S. Oymak, A. Jalali, M. Fazel, and B. Hassibi, "Noisy Estimation of Simultaneously Structured Models: Limitations of Convex Relaxation", available at http://faculty.washington.edu/mfazel/SimultaneousEstimation.pdf.