Identifiability Scaling Laws in Bilinear Inverse Problems

Sunav Choudhary, *Student Member, IEEE,* and Urbashi Mitra, *Fellow, IEEE*

Abstract

A number of ill-posed inverse problems in signal processing, like blind deconvolution, matrix factorization, dictionary learning and blind source separation share the common characteristic of being bilinear inverse problems (BIPs), *i.e.* the observation model is a function of two variables and conditioned on one variable being known, the observation is a linear function of the other variable. A key issue that arises for such inverse problems is that of identifiability, *i.e.* whether the observation is sufficient to unambiguously determine the pair of inputs that generated the observation. Identifiability is a key concern for applications like blind equalization in wireless communications and data mining in machine learning. Herein, a unifying and flexible approach to identifiability analysis for general *conic prior* constrained BIPs is presented, exploiting a connection to low-rank matrix recovery via 'lifting'. We develop deterministic identifiability conditions on the input signals and examine their satisfiability in practice for three classes of signal distributions, *viz.* dependent but uncorrelated, independent Gaussian, and independent Bernoulli. In each case, scaling laws are developed that trade-off probability of robust identifiability with the complexity of the rank two null space. An added appeal of our approach is that the rank two null space can be partly or fully characterized for many bilinear problems of interest (*e.g.* blind deconvolution). We present numerical experiments involving variations on the blind deconvolution problem that exploit a characterization of the rank two null space and demonstrate that the scaling laws offer good estimates of identifiability.

Index Terms

Bilinear inverse problems, blind deconvolution, identifiability, rank one matrix recovery

I. Introduction

We examine the problem of identifiability in bilinear inverse problems (BIPs), *i.e.* input signal pair recovery for systems where the output is a bilinear function of two unknown inputs. Important practical examples of BIPs include blind deconvolution [3], blind source separation [4] and dictionary learning [5] in signal processing, matrix factorization in machine learning [6], blind equalization in wireless communications [7], *etc.* Of particular interest are signal recovery problems from *under-determined* systems of measurement where additional structure is needed in order to ensure recovery, and the observation model is *non-linear* in the parametrization of the problem.

Consider a discrete-time blind linear deconvolution problem. Let $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}$ and $\boldsymbol{y} \in \mathcal{D}_{\boldsymbol{y}}$ be respectively m and n dimensional vectors from domains $\mathcal{D}_{\boldsymbol{x}} \subseteq \mathbb{R}^m$ and $\mathcal{D}_{\boldsymbol{y}} \subseteq \mathbb{R}^n$, and suppose that the noise free linear convolution of $\boldsymbol{x}$ and $\boldsymbol{y}$ is observed as $\boldsymbol{z}$. Then the blind linear deconvolution problem can be represented as the following feasibility problem.

$$\begin{aligned} \text{find} \quad & (\boldsymbol{x}, \boldsymbol{y}) \\ \text{subject to} \quad & \boldsymbol{x} \star \boldsymbol{y} = \boldsymbol{z}, \\ & \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}}, \boldsymbol{y} \in \mathcal{D}_{\boldsymbol{y}}. \end{aligned} \tag{P$_1$}$$

We draw the reader's attention to the observation/measurement model $\boldsymbol{z} = \boldsymbol{x} \star \boldsymbol{y}$. Notice that if either $\boldsymbol{x}$ or $\boldsymbol{y}$ was a fixed and known quantity, then we would have an observation model that is linear in the other variable. However, when both $\boldsymbol{x}$ and $\boldsymbol{y}$ are unknown variables, then the linear convolution measurement model $\boldsymbol{z} = \boldsymbol{x} \star \boldsymbol{y}$ is no longer linear in the variable pair $(\boldsymbol{x}, \boldsymbol{y})$. Such a structural characteristic is referred to as a *bilinear* measurement structure (formally defined in Section II). The blind linear deconvolution problem (P$_1$) is the resulting inverse problem. Such inverse problems arising from a bilinear measurement structure shall be referred to as *bilinear inverse problems* (formally defined in Section II).

A key issue in many under-determined inverse problems is that of *identifiability*: "Does a unique solution exist that satisfies the given observations?" Identifiability (and signal reconstruction) for *linear* inverse problems with sparsity and low-rank structures has received considerable attention in the context of compressed sensing and low-rank matrix recovery, respectively, and are now quite well understood [8]. In a nutshell, both compressed sensing and low-rank matrix recovery theories guarantee that the unknown sparse/low-rank signal can be *unambiguously* reconstructed from relatively few properly designed linear measurements using algorithms with runtime growing polynomially in the signal dimension. For non-linear inverse problems (including BIPs), however, characterization of identifiability (and signal reconstruction) still remains largely open. To illustrate

This work has been funded in part by the following grants and organizations: ONR N00014-09-1-0700, NSF CNS-0832186 and NSF CCF-1117896. Parts of this paper were presented at the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Vancouver, Canada, May 26-31, 2013 [1] and at the 47th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, California, Nov. 3-6, 2013 [2].

S. Choudhary and U. Mitra are with the Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles CA 90089, USA (email: sunavcho@usc.edu, ubli@usc.edu)

that analyzing identifiability is nontrivial, we present a simple example. Consider the blind linear deconvolution problem represented by Problem (P_1). Suppose that we have the observation $\boldsymbol{z} = (1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,0)^{\mathrm{T}} \in \mathbb{R}^{17}$ with $\mathcal{D}_{\boldsymbol{x}} = \mathbb{R}^7$ and $\mathcal{D}_{\boldsymbol{y}} = \mathbb{R}^{11}$. It is not difficult to verify that both

$$\boldsymbol{x} = (1,0,0,0,1,0,0)^{\mathrm{T}},\ \boldsymbol{y} = (1,0,1,0,0,0,0,0,1,0,1)^{\mathrm{T}} \tag{1a}$$

and,

$$\boldsymbol{x} = (1,0,1,0,1,0,1)^{\mathrm{T}},\ \boldsymbol{y} = (1,0,0,0,0,0,0,0,1,0,0)^{\mathrm{T}} \tag{1b}$$

are valid solutions to Problem (P_1). Furthermore, it is not immediately obvious as to what structural constraints would disambiguate between the above two solutions. We have showed identifiability and constructed fast recovery algorithms in a previous work [9] when $\boldsymbol{x}$ (possibly sparse) is in the non-negative orthant (modulo global sign flip), whereas we show negative results for the more general sparse (with respect to the canonical basis) blind deconvolution problem in [10], [11].

A. Contributions

1) We cast conic prior constrained BIPs as low-rank matrix recovery problems, establish the validity of the 'lifting' procedure (Section II-C) and develop deterministic sufficient conditions for identifiability (Section III-B) while bridging the gap to necessary conditions in a special case. Our characterization agrees with the intuition that identifiability subject to priors should depend on the joint geometry of the signal space and the bilinear map. Our results are geared towards bilinear maps that admit a nontrivial rank two null space, as is the case with many important BIPs like blind deconvolution.
2) We develop trade-offs between probability of identifiability of a random instance and the complexity of the rank two null space of the lifted bilinear map under three classes of signal ensembles, *viz.* dependent but uncorrelated, independent Gaussian, and independent Bernoulli (Section III-D). Specifically, we demonstrate that instance identifiability can be characterized by the complexity of restricted rank two null space, measured by the covering number of the set $\{(\mathcal{C}(\boldsymbol{X}), \mathcal{R}(\boldsymbol{X})) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}\}$, where $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{R}(\boldsymbol{X})$ denote, respectively, the column and row spaces of the matrix $\boldsymbol{X}$ and $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ denotes the rank two null space of the lifted bilinear map $\mathscr{S}(\cdot)$ restricted by the prior on the signal set to $\mathcal{M}$. To the best of our knowledge, this gives new structural results solely based on the bilinear measurement model and is thus applicable to general BIPs.
3) We demonstrate that the rank two null space of the lifted bilinear map can be partly characterized in at least one important case (blind deconvolution), and conjecture that the same should be possible for other bilinear maps of interest (dictionary learning, blind source separation, *etc.*). Based on this characterization, we present numerical simulations for selected variations on the blind deconvolution problem to demonstrate the tightness of our scaling laws (Section V).

B. Related Work

Our treatment of BIPs draws on several different ideas. We employ 'lifting' from optimization [12] which enables the creation of good relaxations for intractable optimization problems. This can come at the expense of an increase in the ambient dimension of the optimization variables. Lifting was used in [13] for analyzing the phase retrieval problem and in [14] for the analysis of blind circular deconvolution. We employ lifting in the same spirit as [13], [14] but our *goals are different*. Firstly, we deal with general BIPs which include the linear convolution model of [15], the circular convolution model of [14], [16] and the compressed bilinear observation model of [17] as special cases. Secondly, we focus solely on identifiability (as opposed to recoverability by convex optimization [13], [14]) enabling far milder assumptions on the distribution of the input signals.

After lifting, we have a rank one matrix recovery problem, subject to inherited conic constraints. While encouraging results have been shown for low-rank matrix recovery using the nuclear norm heuristic [18], quite stringent incoherence assumptions are needed between the sampling operator and the true matrix. Furthermore, the results do not generalize to an analysis of identifiability when the sampling operator admits rank two matrices in its null space. We are able to relax the incoherence assumptions in special cases for analyzing identifiability and also consider sampling operators with a non-trivial rank two null space. Since the works [19]–[21] can be interpreted as solving BIPs with the lifted map drawn from a Gaussian random ensemble, thus leading to a trivial rank two null space with high probability, the results therein are not directly comparable to our results.

In [14], a recoverability analysis for the blind circular deconvolution problem is undertaken, but the knowledge of the sparsity pattern of one input signal is needed. Taking our Problem (P_1) as an example, [14] assumes $\mathcal{D}_{\boldsymbol{x}} = \mathcal{C}(\boldsymbol{B})$ and $\mathcal{D}_{\boldsymbol{y}} = \mathcal{C}(\boldsymbol{C})$ for some *tall* deterministic matrix $\boldsymbol{B}$ and a *tall* Gaussian random matrix $\boldsymbol{C}$, where for any matrix $\boldsymbol{X}$, $\mathcal{C}(\boldsymbol{X})$ denotes the column space of $\boldsymbol{X}$. In contrast, we shall make the less stringent assumption on $\boldsymbol{x}$ and $\boldsymbol{y}$ and show that *identifiability* holds with high probability in the presence of rank two matrices in the null space of the lifted linear operator (sampling operator).

A closely related (but different) problem is that of retrieving the phase of a signal from the magnitude of its Fourier coefficients (the Fourier phase retrieval problem). This is equivalent to recovering the signal given its auto correlation function [22]. In terms of our example blind deconvolution problem (P_1), phase retrieval is equivalent to having the additional constraints $\mathcal{D}_{\boldsymbol{x}} = \mathcal{D}_{\boldsymbol{y}}$ and $\boldsymbol{y}$ being the time reversed version of $\boldsymbol{x}$. While the Fourier phase retrieval problem may seem superficially similar to the blind deconvolution problem, there are major differences between the two, so much as to ensure identifiability and efficient

recoverability for the sparsity regularized (in the canonical basis) version of the former [23], while one can explicitly show *unidentifiability* for the sparsity regularized (in the canonical basis) version of the latter [10], [11] (even with oracle knowledge of the supports of both signals). The difference arises because the Fourier phase retrieval problem is a (non-convex) quadratic inverse problem rather than a BIP, and it satisfies additional properties (constant trace of the lifted variable) which make it better conditioned for efficient recovery algorithms [24].

For the dictionary learning problem, an identifiability analysis is developed in [25] leveraging results from [26] on matrix factorization for sparse dictionary learning using the ℓ_1 norm and ℓ_p quasi-norm for $0 < p < 1$. More recently, exact recoverability of over-complete dictionaries from training samples (only polynomially large in the dimensions of the dictionary) has been proved in [27] assuming sparse (but unknown) coefficient matrix. While every BIP can be recast as a dictionary learning problem in principle, such a transformation would result in additional structural constraints on the dictionary that may or may not be trivial to incorporate in the existing analyses. This is especially true for bilinear maps over vector pairs. In contrast, we develop our methods to specifically target bilinear maps over vector pairs (*e.g.* convolution map) and thus obtain definitive results where the dictionary learning based formulations would most likely fail.

Some identifiability results for blind deconvolution are summarized in [28], but the treatment therein is inflexible to the inclusion of side information about the input signals. Identifiability for non-negative matrix factorization was examined in [6] exploiting geometric properties of the non-negative orthant. Although our results can be easily visualized in terms of geometry, they can also be stated purely in terms of linear algebra (Theorem 2). Identifiability results for low-rank matrix completion [29], [30] are provided in [31] via algebraic and combinatorial conditions using graph theoretic tools, but there is no straightforward way to extend these results to more general lifted linear operators like the convolution map. Overall, to the best of our knowledge, a unified flexible treatment of identifiability in BIPs has not been developed till date. In this paper, we present such a framework incorporating conic constraints on the input signals (which includes sparse signals in particular).

C. Organization, Reading Guide and Notation

The remainder of the paper is organized as follows. The first half of Section II formally introduces BIPs and a working definition of identifiability. Section II-C describes the lifting technique to reformulate BIPs as rank one matrix recovery problems, and characterizes the validity of the technique. Section III states our main results on both deterministic and random instance identifiability. Section IV elaborates on the intuitions, ideas, assumptions and subtle implications associated with the results of Section III. Section V is devoted to results of numerical verification and Section VI concludes the paper. Detailed proofs of all the results in the paper appear in the Appendices A-O.

In order to maintain linearity of exposition to the greatest extent possible, we chose to create a separate section (Section IV) for elaborating on intuitions, ideas, assumptions and implications associated with the important results of the paper. Thus, with the exception of Section IV, rest of the paper can be read in a linear fashion. However, we recommend the reader to switch between Sections III and IV as necessary, to better interpret the results presented in Section III.

We state the notational conventions used throughout rest of the paper. All vectors are assumed to be column vectors unless stated otherwise. We shall use lowercase boldface alphabets to denote column vectors (*e.g.* $\boldsymbol{z}$) and uppercase boldface alphabets to denote matrices (*e.g.* $\boldsymbol{A}$). The all zero (respectively all one) vector/matrix shall be denoted by $\mathbf{0}$ (respectively $\mathbf{1}$) and the identity matrix by $\mathbf{I}$. The canonical base matrices for the space of $m \times n$ real matrices will be denoted by $\boldsymbol{E}_{i,j}$ for $1 \leq i \leq m$, $1 \leq j \leq n$ and is defined (element-wise) as

$$(\boldsymbol{E}_{i,j})_{k,l} = \begin{cases} 1, & i = k, j = l, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

For vectors and/or matrices, $(\cdot)^{\mathrm{T}}$, $\mathrm{Tr}(\cdot)$ and $\mathrm{rank}(\cdot)$ respectively denote the transpose, trace and rank of their argument, whenever applicable. Special sets are denoted by uppercase blackboard bold font (*e.g.* $\mathbb{R}$ for real numbers). Other sets are denoted by uppercase calligraphic font (*e.g.* $\mathcal{S}$). Linear operators on matrices are denoted by uppercase script font (*e.g.* $\mathscr{S}$). The set of all matrices of rank at most k in the null space of a linear operator $\mathscr{S}$ will be denoted by $\mathcal{N}(\mathscr{S}, k)$, defined as

$$\mathcal{N}(\mathscr{S}, k) \triangleq \left\{ \boldsymbol{X} \in \mathbb{R}^{m \times n} \;\middle|\; \mathrm{rank}(\boldsymbol{X}) \leq k,\, \mathscr{S}(\boldsymbol{X}) = \mathbf{0} \right\}, \tag{3}$$

and referred to as the 'rank k null space'. For any matrix $\boldsymbol{X}$, we denote the row and column spaces by $\mathcal{R}(\boldsymbol{X})$ and $\mathcal{C}(\boldsymbol{X})$ respectively. The projection matrix onto the column space (respectively row space) of $\boldsymbol{X}$ shall be denoted by $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}$ (respectively $\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}$). For any rank one matrix $\boldsymbol{M}$, an expression of the form $\boldsymbol{M} = \sigma \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ would denote the singular value decomposition of $\boldsymbol{M}$ with vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ each admitting unit ℓ_2-norm. The standard Euclidean inner product on a vector space will be denoted by $\langle \cdot, \cdot \rangle$ and the underlying vector space will be clear from the usage context. All logarithms are with respect to (*w.r.t.*) base e unless specified otherwise. We shall use the $O(h)$, $o(h)$ and $\Theta(h)$ notation to denote order of growth of any function $f\colon \mathbb{R} \to \mathbb{R}$ of $h \in \mathbb{R}$ *w.r.t.* its argument. We have,

$$f(h) = O(h) \iff \lim_{h \to \infty} \frac{f(h)}{h} < \infty, \tag{4a}$$

$$f(h) = o(h) \iff \lim_{h\to\infty} \frac{f(h)}{h} = 0, \tag{4b}$$

$$f(h) = \Theta(h) \iff \lim_{h\to\infty} \frac{f(h)}{h} \in (0, \infty). \tag{4c}$$

II. System Model

This section introduces the bilinear observation model and the associated bilinear inverse problem in Subsection II-A and our working definition of identifiability in Subsection II-B. Subsection II-C describes the equivalent linear inverse problem obtained by lifting and conditions under which the equivalence holds. This equivalence is used to establish all of our identifiability results in Section III.

A. Bilinear Maps and Bilinear Inverse Problems (BIPs)

Definition 1 (Bilinear Map)**.** *A mapping $\boldsymbol{S}\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^q$ is called a bilinear map if $\boldsymbol{S}(\cdot, \boldsymbol{y})\colon \mathbb{R}^m \to \mathbb{R}^q$ is a linear map $\forall \boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{S}(\boldsymbol{x}, \cdot)\colon \mathbb{R}^n \to \mathbb{R}^q$ is a linear map $\forall \boldsymbol{x} \in \mathbb{R}^m$.*

We shall consider the generic bilinear system/measurement model introduced in [1],

$$\boldsymbol{z} = \boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}), \tag{5}$$

where $\boldsymbol{z}$ is the vector of observations, $\boldsymbol{S}\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^q$ is a given bilinear map, and $(\boldsymbol{x}, \boldsymbol{y})$ denotes the pair of unknown signals with a given domain restriction $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}$. We are interested in solving for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ from the noiseless observation $\boldsymbol{z}$ as given by (5). The BIP corresponding to the observation model (5) is represented by the following feasibility problem.

$$\begin{aligned} &\text{find} && (\boldsymbol{x}, \boldsymbol{y}) \\ &\text{subject to} && \boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{z}, \\ &&& (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}. \end{aligned} \tag{P_2}$$

The non-negative matrix factorization problem [6] serves as an illustrative example of such a problem. Let $\boldsymbol{X} \in \mathbb{R}^{m\times k}$ and $\boldsymbol{Y} \in \mathbb{R}^{k\times n}$ be two element-wise non-negative, unknown matrices and suppose that we observe the matrix product $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Y}$ which clearly has a bilinear structure. The non-negative matrix factorization problem is represented by the feasibility problem

$$\begin{aligned} &\text{find} && (\boldsymbol{X}, \boldsymbol{Y}) \\ &\text{subject to} && \boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Y}, \\ &&& \boldsymbol{X} \geq \boldsymbol{0}, \boldsymbol{Y} \geq \boldsymbol{0}. \end{aligned} \tag{P_3}$$

where the expressions $\boldsymbol{X} \geq \boldsymbol{0}$ and $\boldsymbol{Y} \geq \boldsymbol{0}$ constrain the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be elementwise non-negative. The elementwise non-negativity constraints $\boldsymbol{X} \geq \boldsymbol{0}, \boldsymbol{Y} \geq \boldsymbol{0}$ form a domain restriction in Problem (P_3), in the same way as the constraint $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}$ serves to restrict the feasible set in Problem (P_2).

B. Identifiability Definition

Notice that *every* BIP has an inherent scaling ambiguity due to the identity

$$\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{S}\left(\alpha\boldsymbol{x}, \frac{1}{\alpha}\boldsymbol{y}\right), \quad \forall \alpha \neq 0, \tag{6}$$

where $\boldsymbol{S}(\cdot, \cdot)$ represents the bilinear map. Thus, a meaningful definition of identifiability, in the context of BIPs, must disregard this type of scaling ambiguity. This leads us to the following definition of identifiability.

Definition 2 (Identifiability)**.** *A vector pair $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K} \subseteq \mathbb{R}^m \times \mathbb{R}^n$ is identifiable w.r.t. the bilinear map $\boldsymbol{S}\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^q$ if $\forall (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{K} \subseteq \mathbb{R}^m \times \mathbb{R}^n$ satisfying $\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{S}(\boldsymbol{x}', \boldsymbol{y}')$, $\exists \alpha \neq 0$ such that $(\boldsymbol{x}', \boldsymbol{y}') = \left(\alpha\boldsymbol{x}, \frac{1}{\alpha}\boldsymbol{y}\right)$.*

Remark 1. *It is straightforward to see that our definition of identifiability in turn defines an equivalence class of solutions. Thus, we seek to identify the equivalence class induced by the observation $\boldsymbol{z}$ in (5). Later, in Section II-C, we shall 'lift' Problem (P_2) to Problem (P_4) where, every equivalence class in the domain $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}$ of the former problem maps to a single point in the domain $\boldsymbol{W} \in \mathcal{K}'$ of the latter problem.*

Remark 2. *The scaling ambiguity represented by (6) is common to all BIPs and our definition of identifiability (Definition 2) only allows for this kind of ambiguity. There may be other types of ambiguities depending on the specific BIP. For example, the forward system model associated with Problem (P_3) is given by the matrix product operation $\boldsymbol{S}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{Y}$ which shows the following matrix multiplication ambiguity.*

$$\boldsymbol{S}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{S}\left(\boldsymbol{X}\boldsymbol{T}, \boldsymbol{T}^{-1}\boldsymbol{Y}\right) \tag{7}$$

where $\boldsymbol{T}^{-1}$ is the right inverse of $\boldsymbol{T}$. It is possible to define weaker notions of identifiability to allow for this kind of ambiguity. In this paper, we shall not address this question any further and limit ourselves to the stricter notion of identifiability as given by Definition 2.

C. Lifting

While Problem (P_2) is an accurate representation of the class of BIPs, the formulation does not easily lend itself to an identifiability analysis. We next rewrite Problem (P_2) to facilitate analysis, subject to some technical conditions (see Theorem 1 and Corollary 1). The equivalent problem is a matrix rank minimization problem subject to linear equality constraints

$$\begin{aligned} \underset{\boldsymbol{W}}{\text{minimize}} \quad & \operatorname{rank}(\boldsymbol{W}) \\ \text{subject to} \quad & \mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}, \\ & \boldsymbol{W} \in \mathcal{K}', \end{aligned} \tag{P_4}$$

where $\mathcal{K}' \subseteq \mathbb{R}^{m\times n}$ is *any* set satisfying

$$\mathcal{K}' \bigcap \left\{ \boldsymbol{W} \in \mathbb{R}^{m\times n} \;\middle|\; \operatorname{rank}(\boldsymbol{W}) \leq 1 \right\} = \left\{ \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K} \right\}, \tag{8}$$

and $\mathscr{S}\colon \mathbb{R}^{m\times n} \to \mathbb{R}^q$ is a linear operator that can be *deterministically* constructed from the bilinear map $\boldsymbol{S}(\cdot,\cdot)$ with the optimization variable $\boldsymbol{W}$ in Problem (P_4) being related to the optimization variable pair $(\boldsymbol{x},\boldsymbol{y})$ in Problem (P_2) by the relation $\boldsymbol{W} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$. The transformation of Problem (P_2) to Problem (P_4) is an example of 'lifting' and we shall refer to $\mathscr{S}(\cdot)$ as the 'lifted linear operator' *w.r.t.* the bilinear map $\boldsymbol{S}(\cdot,\cdot)$. Other examples on lifting can be found in [13], [14]. Before stating the equivalence results between Problems (P_2) and (P_4) we describe the construction of $\mathscr{S}(\cdot)$ from $\boldsymbol{S}(\cdot,\cdot)$.

Let $\phi_j\colon \mathbb{R}^q \to \mathbb{R}$ be the j^{th} coordinate projection operator of q dimensional vectors to scalars, *i.e.* if $\boldsymbol{z} = (z_1, z_2, \ldots, z_q)$ then $\phi_j(\boldsymbol{z}) = z_j$. Clearly, ϕ_j is a linear operator and hence the composition $\phi_j \circ \boldsymbol{S}\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a bilinear map. As $\boldsymbol{S}$ is a finite dimensional operator, it is a bounded operator, hence by the Riesz Representation Theorem [32], $\exists \boldsymbol{S}_j \in \mathbb{R}^{m\times n}$ such that $\boldsymbol{S}_j$ is the unique linear operator satisfying

$$\phi_j \circ \boldsymbol{S}(\boldsymbol{x},\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{S}_j\boldsymbol{y} \rangle, \quad \forall \boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n, \tag{9}$$

where $\langle \cdot,\cdot \rangle$ denotes an inner product operation in $\mathbb{R}^m$. Using (9), we can convert the bilinear equality constraint in Problem (P_2) into a set of q linear equality constraints as follows:

$$z_j = \phi_j \circ \boldsymbol{S}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}_j\boldsymbol{y} = \left\langle \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}, \boldsymbol{S}_j \right\rangle \tag{10}$$

for each $1 \leq j \leq q$, where the last inner product in (10) is the trace inner product in the space $\mathbb{R}^{m\times n}$ and z_j denotes the j^{th} coordinate of the observation vector $\boldsymbol{z}$. Setting $\boldsymbol{W} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ in (10), the q linear equality constraints in (10) can be compactly represented, using operator notation, by the vector equality constraint $\mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}$, where $\mathscr{S}\colon \mathbb{R}^{m\times n} \to \mathbb{R}^q$ is a linear operator acting on $\boldsymbol{W} \in \mathbb{R}^{m\times n}$. This derivation uniquely specifies $\mathscr{S}(\cdot)$ using the matrices $\boldsymbol{S}_j$, $1 \leq j \leq q$, and we have the identity

$$\mathscr{S}\left(\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\right) = \boldsymbol{S}(\boldsymbol{x},\boldsymbol{y}), \quad \forall (\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^m \times \mathbb{R}^n. \tag{11}$$

For the sake of completeness, we state the definitions of *equivalence* and *feasibility* in the context of optimization problems (Definitions 3 and 4). Thereafter, the connection between Problems (P_2) and (P_4) is described via the statements of Theorem 1 and Corollary 1.

Definition 3 (Equivalence of optimization problems)**.** *Two optimization problems P and Q are said to be equivalent if every solution to P gives a solution to Q and every solution to Q gives a solution to P.*

Definition 4 (Feasibility)**.** *An optimization problem is said to be feasible, if the domain of the optimization variable is non-empty.*

Theorem 1. *Let Problem (P_2) be feasible and let $\mathcal{K}_{opt}$ and $\mathcal{K}'_{opt}$ denote the set of solutions to Problems (P_2) and (P_4), respectively. Then the following are true.*

1) *Problem (P_4) is feasible with solution(s) of rank at most one.*
2) $\mathcal{K}'_{opt} \subseteq \left\{ \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}_{opt} \right\}$.
3) $\mathcal{K}'_{opt} = \left\{ \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}_{opt} \right\}$ *if and only if* $\{\boldsymbol{0}\} \subsetneq \left\{ \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}_{opt} \right\}$ *does not hold.*

Proof: Appendix A. ∎

Notice that $\mathcal{K}_{\text{opt}}$ and $\mathcal{K}'_{\text{opt}}$ in Theorem 1 depend on the observation vector $\boldsymbol{z}$, so that the statements of Theorem 1 have a hidden dependence on $\boldsymbol{z}$. Since the observation vector $\boldsymbol{z}$ is a function of the input signal pair $(\boldsymbol{x},\boldsymbol{y})$ it is desirable to have statements analogous to Theorem 1 that do not depend on the observation vector $\boldsymbol{z}$. This is the purpose of Corollary 1 below which makes use of $\mathcal{N}(\mathscr{S},1)$, the rank one null space of the lifted operator $\mathscr{S}(\cdot)$ (see (3)).

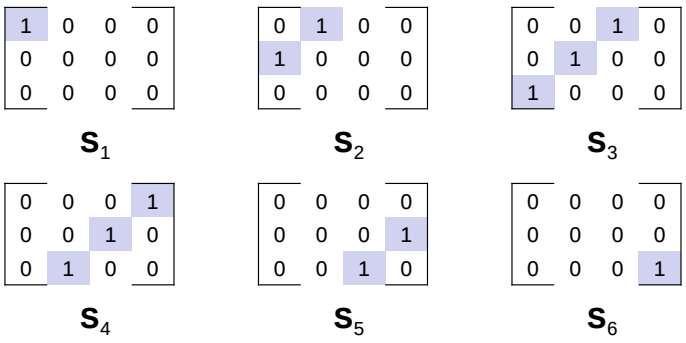

Fig. 1. Lifted matrices $\boldsymbol{S}_k \in \mathbb{R}^{m\times n}$ for linear convolution map with $m = 3$, $n = 4$, $q = m + n - 1 = 6$ and $1 \leq k \leq q$.

Corollary 1. *Let Problem* (P_2) *be feasible and let* $\mathcal{K}_{opt}(\boldsymbol{z})$ *and* $\mathcal{K}'_{opt}(\boldsymbol{z})$ *respectively denote the set of optimal solutions to Problems* (P_2) *and* (P_4) *for a given observation vector* $\boldsymbol{z}$*. Problems* (P_2) *and* (P_4) *are equivalent, i.e.* $\mathcal{K}'_{opt}(\boldsymbol{z}) = \left\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}_{opt}(\boldsymbol{z})\right\}$*, for every* $\boldsymbol{z} \in \{\boldsymbol{S}(\boldsymbol{x},\boldsymbol{y}) \mid (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}\}$ *if and only if* $\{\boldsymbol{0}\} \subsetneq \mathcal{K}' \bigcap \mathcal{N}(\mathscr{S}, 1)$ *does not hold.*

Proof: Appendix B. ∎

Remark 3. *The statements of Theorem 1 and Corollary 1 are needed to establish the validity of lifting for general BIPs with* $\mathcal{N}(\mathscr{S}, 1) \neq \{\boldsymbol{0}\}$*. In case* $\mathcal{N}(\mathscr{S}, 1) = \{\boldsymbol{0}\}$ *(e.g. blind deconvolution), Corollary 1 immediately implies that lifting is valid.*

Remark 4. *Notice that lifting Problem* (P_2) *to Problem* (P_4) *allows us some freedom in the choice of the set* $\mathcal{K}'$*. Also, we have the additional side information that the optimal solution to Problem* (P_4) *is a rank one matrix. These factors could be potentially helpful to develop tight and tractable relaxations to Problem* (P_4)*, that work better than the simple nuclear norm heuristic [33] (e.g. see [27]). We do not pursue this question here.*

This transformation from Problem (P_2) to Problem (P_4) gives us several advantages,

1) Problem (P_4) has linear equality constraints as opposed to the bilinear equality constraints of Problem (P_2). The former is much easier to handle from an optimization as well as algorithmic perspective than the latter.
2) Convex relaxation for the nonconvex rank constraint in Problem (P_4) is well known [33], which is an important requirement from an algorithmic perspective. In contrast, convex relaxation for a generic bilinear constraint is not known.
3) The bilinear map is completely determined by the set of matrices $\boldsymbol{S}_j$ and is separated from the variable $\boldsymbol{W}$ in Problem (P_4). Thus, Problem (P_4) can be used to study generic BIPs. Fig. 1 illustrates a toy example involving the linear convolution map.
4) For every BIP there is an inherent scaling ambiguity (see (6)) associated with the bilinear constraint. However, in Problem (P_4), this scaling ambiguity has been taken care of implicitly when $\boldsymbol{W} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ is the variable to be determined. Clearly, $\boldsymbol{W}$ is unaffected by the type of scaling ambiguity described in (6). Norm constraints on $\boldsymbol{x}$ or $\boldsymbol{y}$ can be used to recover $\boldsymbol{x}$ and $\boldsymbol{y}$ from $\boldsymbol{W}$ but these constraints do not affect Problem (P_4).
5) If $\boldsymbol{x}$ and/or $\boldsymbol{y}$ are sparse in some known dictionary (possibly over-complete) then they can be absorbed into the mapping matrices $\boldsymbol{S}_j$ without altering the structure of Problem (P_4). Indeed, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are dictionaries such that $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\beta}$ and $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\gamma}$ then we have
$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{S}_j\boldsymbol{y} = \boldsymbol{\beta}^{\mathrm{T}}\big(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_j\boldsymbol{B}\big)\boldsymbol{\gamma} = \big\langle \boldsymbol{\beta}\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_j\boldsymbol{B}\big\rangle \tag{12}$$
for each $1 \leq j \leq q$. It is clear that Problem (P_4) can be rewritten with $\boldsymbol{W} = \boldsymbol{\beta}\boldsymbol{\gamma}^{\mathrm{T}}$ as the optimization variable (with a corresponding modification to $\mathcal{K}'$), and comparing (12) and (10) we see that the matrix $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_j\boldsymbol{B}$ can be designated to play the same role in the *rewritten* Problem (P_4) as $\boldsymbol{S}_j$ played in the original Problem (P_4). Thus, without loss of generality, we can consider Problem (P_4) to be our lifted problem that retains all available prior information from Problem (P_2) (assuming that the equivalence conditions in Corollary 1 are satisfied).

III. Identifiability Results

We state our main results in this section starting with deterministic characterizations of identifiability in Subsections III-A and III-B that are simple to state but computationally hard to check for a given BIP. Subsequently, in Subsection III-D we investigate whether identifiability holds for most inputs if the input is drawn from some distribution over the domain.

Since we have some freedom of choice in the selection of the set $\mathcal{K}'$ according to Remark 4, we will work with an arbitrary $\mathcal{K}'$ satisfying (8). The extreme cases of $\mathcal{K}' = \left\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{K}\right\}$ and $\mathcal{K}' = \mathbb{R}^{m\times n}$ will sometimes be used for examples and to build intuition. Also, for some of the results, we have converse statements only for one of the extreme cases. We shall use the set $\mathcal{M}$ to denote the difference $\mathcal{K}' - \mathcal{K}'$, defined as
$$\mathcal{M} = \mathcal{K}' - \mathcal{K}' \triangleq \{\boldsymbol{X}_1 - \boldsymbol{X}_2 \mid \boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{K}'\}. \tag{13}$$

A. Universal Identifiability

As a straightforward consequence of lifting, we have the following necessary and sufficient condition for Problem (P_4) to succeed for all values of the observation $\boldsymbol{z} = \boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y})$.

Proposition 1. *Let* $\mathcal{K}' = \left\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\right\}$. *The solution to Problem* (P_4) *will be correct for every observation* $\boldsymbol{z} = \boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y})$ *if and only if* $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$.

Proof: Appendix C. ∎

Remark 5. *Notice that the "only if" part of Proposition 1 requires uniqueness of an observation* $\boldsymbol{z}$ *that is valid for Problem* (P_2) *as well and not just for Problem* (P_4). *The latter could have observations that arise because of the freedom in the choice of* $\mathcal{K}'$, *but those may not be valid for the former. As a result, the conclusion of the "only if" part of Proposition 1 is somewhat weaker in that it does not imply* $\mathcal{N}(\mathscr{S}, 2) = \{\boldsymbol{0}\}$.

When $\mathcal{K} = \mathbb{R}^m \times \mathbb{R}^n$, $\mathcal{M}$ represents the set of all rank two matrices in $\mathbb{R}^{m \times n}$ so that Proposition 1 reduces to the more familiar result: $\mathcal{N}(\mathscr{S}, 2) = \{\boldsymbol{0}\}$ is necessary and sufficient for the action of the linear operator $\mathscr{S}$ to be invertible on the set of all rank one matrices, where the inversion of the action of $\mathscr{S}$ is achieved as the solution to Problem (P_4).

While the characterization of $\mathcal{N}(\mathscr{S}, 2)$ for arbitrary linear operators $\mathscr{S}(\cdot)$ is challenging, it has been shown that if $\mathscr{S}(\cdot)$ is picked as a realization from some desirable distribution then $\mathcal{N}(\mathscr{S}, 2) = \{\boldsymbol{0}\}$ (implies $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$) is satisfied with high probability. As an example, [19], [20] show that if $\mathscr{S}: \mathbb{R}^{m \times n} \to \mathbb{R}^q$ is picked from a Gaussian random ensemble, then $\mathcal{N}(\mathscr{S}, 2) = \{\boldsymbol{0}\}$ is satisfied with high probability for $q = O(\max(m, n))$.

B. Deterministic Instance Identifiability

When $\mathscr{S}(\cdot)$ is sampled from less desirable distributions, as for matrix completion [29], [30] or matrix recovery for a specific given basis [18], one does *not* have $\mathcal{N}(\mathscr{S}, 2) = \{\boldsymbol{0}\}$ with high probability. To guarantee identifiability (and unique reconstruction) for such realizations of $\mathscr{S}(\cdot)$, significant domain restrictions via the set $\mathcal{K}$ (or $\mathcal{K}'$) are usually needed, so that $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$ and Proposition 1 comes into effect. Unfortunately, for many important BIPs (blind deconvolution, blind source separation, matrix factorization, *etc.*) the lifted linear operator $\mathscr{S}(\cdot)$ *does have* a non-trivial $\mathcal{N}(\mathscr{S}, 2)$ set. This makes identifiability an important issue in practice. Fortunately, we still have $\mathcal{N}(\mathscr{S}, 1) = \{\boldsymbol{0}\}$ in many of these cases so that Corollary 1 implies that lifting is valid. For such maps, we have the following deterministic sufficient condition (Theorem 2) for a rank one matrix $\boldsymbol{M} \in \mathcal{K}' \subseteq \mathbb{R}^{m \times n}$ to be identifiable as a solution of Problem (P_4). Theorem 2 is heavily used for the results in the sequel.

Theorem 2. *Let* $\mathcal{N}(\mathscr{S}, 1) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$ *and* $\boldsymbol{M} = \sigma \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ *be a rank one matrix in* $\mathcal{K}' \subseteq \mathbb{R}^{m \times n}$. *Suppose that for every* $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ *either* $\boldsymbol{u} \notin \mathcal{C}(\boldsymbol{X})$ *or* $\boldsymbol{v} \notin \mathcal{R}(\boldsymbol{X})$ *is true, then given the observation* $\boldsymbol{z} = \mathscr{S}(\boldsymbol{M})$, $\boldsymbol{M}$ *can be successfully recovered by solving Problem* (P_4).

Proof: Appendix D. ∎

Theorem 2 is only a sufficient condition for identifiability. We bridge the gap to the necessary conditions under a special case in Corollary 2 below. We use the notation $\boldsymbol{M} - \mathcal{K}'$ to denote the set $\{\boldsymbol{M} - \boldsymbol{Y} \mid \boldsymbol{Y} \in \mathcal{K}'\}$.

Corollary 2. *Let* $\mathcal{N}(\mathscr{S}, 1) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$ *and* $\boldsymbol{M} = \sigma \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ *be a rank one matrix in* $\mathcal{K}' \subseteq \mathbb{R}^{m \times n}$. *Suppose that every matrix* $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap (\boldsymbol{M} - \mathcal{K}') \setminus \{\boldsymbol{0}\}$ *admits a singular value decomposition with* $\sigma_1(\boldsymbol{X}) = \sigma_2(\boldsymbol{X})$. *Let us denote such a decomposition as* $\boldsymbol{X} = \sigma_* \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \sigma_* \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}}$, *and let* $\boldsymbol{u} = \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2$ *and* $\boldsymbol{v} = \alpha_3 \boldsymbol{v}_1 + \alpha_4 \boldsymbol{v}_2$ *for some* $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}$ *with* $\alpha_1^2 + \alpha_2^2 = \alpha_3^2 + \alpha_4^2 = 1$. *Given the observation* $\boldsymbol{z} = \mathscr{S}(\boldsymbol{M})$, *Problem* ($P_4$) *successfully recovers* $\boldsymbol{M}$ *if and only if for every* $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap (\boldsymbol{M} - \mathcal{K}') \setminus \{\boldsymbol{0}\}$, $\alpha_1 \alpha_3 + \alpha_2 \alpha_4 \leq 0$.

Proof: Appendix E. ∎

Intuitively, Corollary 2 exploits the fact that all nonzero singular values of a matrix are of the same sign. Indeed, (α_1, α_2) (respectively (α_3, α_4)) is an element of the two dimensional space of representation coefficients of $\boldsymbol{u}$ *w.r.t.* $\mathcal{C}(\boldsymbol{X})$ (respectively $\boldsymbol{v}$ *w.r.t.* $\mathcal{R}(\boldsymbol{X})$) with a fixed representation basis. Corollary 2 says that identifiability of $\boldsymbol{M}$ holds if and only if the vectors (α_1, α_2) and (α_3, α_4) do not form an acute angle between them. The assumption of $\sigma_1(\boldsymbol{X}) = \sigma_2(\boldsymbol{X})$ has been made in Corollary 2 for ease of intuition. Although we do not state it here, an analogous result holds for $\sigma_1(\boldsymbol{X}) \neq \sigma_2(\boldsymbol{X})$ with the condition on the inner product $\langle (\alpha_1, \alpha_2), (\alpha_3, \alpha_4) \rangle \leq 0$ replaced by the same condition on a weighted inner product, where the weights depend on the ratio of $\sigma_1(\boldsymbol{X})$ to $\sigma_2(\boldsymbol{X})$.

For arbitrary lifted linear operators $\mathscr{S}(\cdot)$, checking Theorem 2 for a given rank one matrix $\boldsymbol{M}$ is usually hard, unless a simple characterization of $\mathcal{N}(\mathscr{S}, 2)$ or $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ has been provided. It is reasonable to ask "How many rank one matrices $\boldsymbol{M}$ are identifiable?", given any particular lifted linear operator $\mathscr{S}(\cdot)$ and assuming that the rank one matrices $\boldsymbol{M}$ are drawn at random from some distribution. It is highly desirable if most rank one matrices $\boldsymbol{M}$ are identifiable. Before we can show such a result we need to define a random model for the rank one matrix $\boldsymbol{M}$.

C. A Random Rank One Model

We consider $\boldsymbol{M} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ as a random rank one matrix drawn from an ensemble with the following properties:

(A1) $\boldsymbol{x}$ (and $\boldsymbol{y}$) is a zero mean random vector with an identity covariance matrix.

(A2) $\boldsymbol{x}$ and $\boldsymbol{y}$ are mutually independent.

As a practical motivation for this random model, we consider a blind channel estimation problem where the transmitted signal $\boldsymbol{x}$ passes through an unknown linear time invariant channel impulse response $\boldsymbol{y}$. In the absence of measurement noise, the observed signal at the receiver would be the linear convolution $\boldsymbol{z} = \boldsymbol{x} \star \boldsymbol{y}$, which is a bilinear map. A practical modeling choice puts the channel realization $\boldsymbol{y}$ statistically independent of the transmitted signal $\boldsymbol{x}$. Furthermore, if channel phase is rapidly varying, then the sign of each entry for $\boldsymbol{y}$ is equally likely to be positive or negative with resultant mean as zero. The transmitted signal $\boldsymbol{x}$ can be assumed to be zero mean with independent and identically distributed entries (and thus identical variance per entry) under Binary-Phase-Shift-Keying and other balanced Phase-Shift-Keying modulation schemes. The assumption of equal variance per tap is somewhat idealistic for channel $\boldsymbol{y}$, but strictly speaking, this requirement is *not* absolutely necessary for our identifiability results.

1) Dependent Entries: First, we consider the case when the elements of $\boldsymbol{x}$ (respectively $\boldsymbol{y}$) are *not independent*. We shall be interested in the following two possible properties of $\boldsymbol{x}$ and $\boldsymbol{y}$:

(A3) The distribution of $\boldsymbol{x}$ (respectively, $\boldsymbol{y}$) factors into a product of marginal distributions of $\|\boldsymbol{x}\|_2$ and $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ (respectively, $\|\boldsymbol{y}\|_2$ and $\boldsymbol{y}/\|\boldsymbol{y}\|_2$).

(A4) $\exists r > 0$ such that $\|\boldsymbol{x}\|_2 \geq r$ (respectively $\|\boldsymbol{y}\|_2 \geq r$) a.s.

We state the following technical lemmas that will be needed in the proofs of Theorem 3 and Corollary 3. Lemma 2 is mainly useful when the assumption (A3) cannot be satisfied but one needs bounds that closely resemble that of Lemma 1. These lemmas allow us to upper bound the probability that $\boldsymbol{x}$ (respectively $\boldsymbol{y}$) is close to one of the key subspaces in Theorem 2, *i.e.* $\mathcal{C}(\boldsymbol{X})$ (respectively $\mathcal{R}(\boldsymbol{X})$) where $\boldsymbol{X}$ is in the appropriately constrained subset of $\mathcal{N}(\mathscr{S}, 2)$.

Lemma 1. *Given any $m \times n$ real matrix $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ and a constant $\delta \in (0, 1)$, a rank one random matrix $\boldsymbol{M} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} = \sigma \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ satisfying assumptions (A1)-(A3) also satisfies,*

$$\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\right\|_2^2 \geq 1 - \delta\right) \leq \frac{2}{m(1-\delta)} \tag{14a}$$

and,

$$\Pr\left(\left\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\right\|_2^2 \geq 1 - \delta\right) \leq \frac{2}{n(1-\delta)}. \tag{14b}$$

Proof: Appendix F. ∎

Lemma 2. *Given any $m \times n$ real matrix $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ and a constant $\delta \in (0, 1)$, a rank one random matrix $\boldsymbol{M} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} = \sigma \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ satisfying assumptions (A1)-(A2), with $\boldsymbol{x}$ (respectively $\boldsymbol{y}$) satisfying (A4) for a constant $r = r_{\boldsymbol{x}}$ (respectively $r = r_{\boldsymbol{y}}$), also satisfies,*

$$\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\right\|_2^2 \geq 1 - \delta\right) \leq \frac{2}{r_{\boldsymbol{x}}^2(1-\delta)} \tag{15a}$$

and,

$$\Pr\left(\left\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\right\|_2^2 \geq 1 - \delta\right) \leq \frac{2}{r_{\boldsymbol{y}}^2(1-\delta)}. \tag{15b}$$

Proof: Appendix G. ∎

Remark 6. *Lemma 2 will give non-trivial bounds if $r_{\boldsymbol{x}}$ (respectively $r_{\boldsymbol{y}}$) go to ∞ fast enough as m (respectively n) goes to ∞, and this growth rate could be slower than $\Theta(\sqrt{m})$ (respectively $\Theta(\sqrt{n})$).*

An example where Lemma 2 is applicable but Lemma 1 is not, can be constructed as follows. As before, let $\boldsymbol{y}$ represent a channel impulse response independent of $\boldsymbol{x}$, so that (A2) is satisfied. Let $\boldsymbol{x}$ represent a coded data stream under Pulse-Amplitude-Modulation such that $\|\boldsymbol{x}\|_2 \in \left\{\sqrt{m/3}, \sqrt{2m/3}\right\}$ with equal probability, $\mathrm{E}[\boldsymbol{x}] = \boldsymbol{0}$ and x_m is coded as a function of $\|\boldsymbol{x}\|_2$ yielding the following conditional correlation matrices:

$$\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \,\middle|\, \|\boldsymbol{x}\|_2 = \sqrt{\frac{m}{3}}\right] = \frac{1}{3}\mathbf{I} + \frac{1}{6}(\boldsymbol{E}_{1,m} + \boldsymbol{E}_{m,1}) \tag{16a}$$

and,

$$\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \,\middle|\, \|\boldsymbol{x}\|_2 = \sqrt{\frac{2m}{3}}\right] = \frac{2}{3}\mathbf{I} - \frac{1}{6}(\boldsymbol{E}_{1,m} + \boldsymbol{E}_{m,1}) \tag{16b}$$

where $\boldsymbol{E}_{i,j} \in \mathbb{R}^{m\times m}$ is the matrix with elements given by

$$(\boldsymbol{E}_{i,j})_{k,l} = \begin{cases} 1, & i=k, j=l, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

for every $1 \leq i,j \leq m$. The expressions in (16) clearly imply that $\|\boldsymbol{x}\|_2$ and $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ are dependent so that (A3) does not hold. Nonetheless, by construction, we have

$$\Pr\left(\|\boldsymbol{x}\|_2 = \sqrt{\frac{m}{3}}\right) = \Pr\left(\|\boldsymbol{x}\|_2 = \sqrt{\frac{2m}{3}}\right) = \frac{1}{2}, \tag{18}$$

so that (16) implies $\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] = \mathbf{I}$, thus satisfying (A1). Also, $\|\boldsymbol{x}\|_2 \geq \sqrt{m/3}$ a.s. so that (A4) is satisfied. Thus, Lemma 2 is applicable with $r_{\boldsymbol{x}} = \sqrt{m/3}$.

2) Independent Entries: While Lemma 1 provides useful bounds, it does not suffice for many problems where $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ is large. We can get much stronger bounds than Lemma 1 if the elements of vector $\boldsymbol{x}$ (respectively $\boldsymbol{y}$) come from independent distributions, by utilizing the *concentration of measure* phenomenon [34]. We shall consider the standard Gaussian and the symmetric Bernoulli distributions, and sharpen the bounds of Lemma 1 in the two technical lemmas to follow. Note that a zero mean independent and identically distributed assumption on the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ already implies the assumptions (A1)-(A3). The bounds of Lemmas 4 and 3 have an interpretation similar to the *restricted isometry property* [35] and are used in the proofs for Theorems 4 and 5, respectively. We retain the assumption $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M} = \{\mathbf{0}\}$ from Theorem 2 and follow the convention that a random variable Z has a symmetric Bernoulli distribution if $\Pr(Z=+1) = \Pr(Z=-1) = 1/2$.

Lemma 3. *Let $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M} = \{\mathbf{0}\}$. Given any $m\times n$ real matrix $\boldsymbol{X} \in \mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}$ and a constant $\delta\in(0,1)$, a random vector $\boldsymbol{x}\in\mathbb{R}^m$ with each element drawn independently from a standard normal distribution satisfies*

$$\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 \geq (1-\delta)\|\boldsymbol{x}\|_2^2\right) \leq \exp\left[-m\log\frac{1}{\sqrt{\delta}} + 2\log m - \frac{2}{m} + 2 - \log\frac{2\delta}{1-\delta}\right]. \tag{19}$$

Proof: Appendix J. ∎

Lemma 4. *Let $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M} = \{\mathbf{0}\}$. Given any $m\times n$ real matrix $\boldsymbol{X} \in \mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}$ and a constant $\delta\in(0,1)$, a random vector $\boldsymbol{x}\in\mathbb{R}^m$ with each element drawn independently from a symmetric Bernoulli distribution satisfies*

$$\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 \geq (1-\delta)\|\boldsymbol{x}\|_2^2\right) \leq \exp\left[-\frac{m(1-\delta)}{4} + \log 4\right]. \tag{20}$$

Proof: Appendix K. ∎

Section IV-F2 provides additional remarks on Lemma 4.

D. Random Instance Identifiability

We first consider the special case where the size of the set $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ is small *w.r.t.* mn, in Section III-D1. We use the same intuition in Section III-D2 to appropriately partition the set $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ when its size is large (possibly infinite) with respect to $m+n$.

1) Small Complexity of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$: It is intuitive to expect that the number of rank one matrices $\boldsymbol{M}$ that are identifiable as optimal solutions to Problem (P_4) should depend inversely on the *size/complexity* of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$. Below, we shall make this notion precise. We shall do so by lower bounding the probability of satisfaction of the sufficient conditions in Theorem 2.

Theorem 3. *Let $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M} = \{\mathbf{0}\}$ and $\boldsymbol{M} = \sigma\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}} \in \mathcal{K}' \subseteq \mathbb{R}^{m\times n}$ be a rank one random matrix satisfying assumptions (A1)-(A3). Suppose that the set $\{(\mathcal{C}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{X})) \mid \boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}\}$ is finite with cardinality $f_{\mathscr{S},\mathcal{M}}(m,n)$. For any constant $\delta\in(0,1)$, the sufficient conditions of Theorem 2 are satisfied with probability greater than $\left(1 - \frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{mn(1-\delta)}\right)$.*

Proof: We describe the basic idea behind the proof and defer the full proof to Appendix H. The proof consists of the following important steps.

(a) We fix the matrix $\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}$ and then relax the "hard" event of subspace membership $\{\boldsymbol{u}\in\mathcal{C}(\boldsymbol{X})\}$ to the "soft" event of being close to the subspace in ℓ_2-norm $\left\{\left\|\boldsymbol{u}-\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\right\|_2^2 \leq \delta\right\}$. This "soft" event describes a body of nonzero volume in $\mathbb{R}^m$. A similar argument holds for the vector $\boldsymbol{v}$ as well.

(b) Next, the volumes (probabilities) of both these bodies (events) is computed individually and utilizing independence between realizations of $\boldsymbol{u}$ and $\boldsymbol{v}$, the probability of the intersection of these events is easily computed. The bounds of Lemma 1 are used in this step.

(c) Lastly, we employ a union bound over the set of valid matrices $\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}$ to make our results universal in nature.

■

Sections IV-A, IV-B and IV-C provide additional remarks on Theorem 3.

In Theorem 3, we can drive the probability of identifiability $\left(1-\frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{mn(1-\delta)}\right)$ arbitrarily close to one by increasing m and/or n provided that $f_{\mathscr{S},\mathcal{M}}(m,n)$ grows as $o(mn)$. For many important BIPs (blind deconvolution, blind source separation, matrix factorization, *etc.*) this growth rate requirement on $f_{\mathscr{S},\mathcal{M}}(m,n)$ is too pessimistic. Tighter versions of Theorem 3, with more optimistic growth rate requirements on $f_{\mathscr{S},\mathcal{M}}(m,n)$, are possible if the assumptions of Lemma 3 or 4 are satisfied. This is the content of Theorems 4 and 5 described in Section III-D2.

We provide a corollary to Theorem 3 when assumption (A3) does not hold so that Lemma 1 is inapplicable. The result uses Lemma 2 in place of Lemma 1 for the proof. The bound is asymptotically useful if $f_{\mathscr{S},\mathcal{M}}(m,n)$ grows as $o\big(r_{\boldsymbol{x}}^2(m)r_{\boldsymbol{y}}^2(n)\big)$.

Corollary 3. *Let $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M}=\{\mathbf{0}\}$ and $\boldsymbol{M}=\sigma\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\in\mathcal{K}'\subseteq\mathbb{R}^{m\times n}$ be a rank one random matrix satisfying assumptions (A1)-(A2) with $\boldsymbol{x}$ (respectively $\boldsymbol{y}$) satisfying (A4) for a constant $r=r_{\boldsymbol{x}}(m)$ (respectively $r=r_{\boldsymbol{y}}(n)$). Suppose that the set $\{(\mathcal{C}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{X}))\mid\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}\}$ is finite with cardinality $f_{\mathscr{S},\mathcal{M}}(m,n)$. For any constant $\delta\in(0,1)$, the sufficient conditions of Theorem 2 are satisfied with probability greater than $\left(1-\dfrac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{r_{\boldsymbol{x}}^2(m)r_{\boldsymbol{y}}^2(n)(1-\delta)}\right)$.*

Proof: Appendix I. ■

2) Large/Infinite Complexity of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$: When the complexity of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ is infinite or exponentially large in $m+n$, the bounds of Section III-D1 become trivially true for large enough m or n. We investigate an alternative bounding technique for this situation using covering numbers. Intuitively speaking, covering numbers measure the size of discretized versions of uncountable sets. The advantage of using such an approach is that the results are not contingent upon the exact geometry of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$. Thus, like Theorem 3, the technique and subsequent results are applicable to every bilinear map. We shall see that to arrive at any sensible results, we will need to use the tighter estimates given by Lemmas 3 and 4 that are only possible when our signals $\boldsymbol{x}$ and $\boldsymbol{y}$ are component-wise independent.

Definition 5 (Covering Number and Metric Entropy [36])**.** *For any two sets $\mathcal{B},\mathcal{D}\subseteq\mathbb{R}^n$, the minimum number of translates of $\mathcal{B}$ needed to cover $\mathcal{D}$ is called the covering number of $\mathcal{D}$ w.r.t. $\mathcal{B}$ and is denoted by $N(\mathcal{D},\mathcal{B})$. The quantity $\log N(\mathcal{D},\mathcal{B})$ is known as the metric entropy of $\mathcal{D}$ w.r.t. $\mathcal{B}$.*

It is known that if $\mathcal{D}\subseteq\mathbb{R}^n$ is a bounded convex body that is symmetric about the origin, and we let $\mathcal{B}=\epsilon\mathcal{D}\triangleq\{\epsilon\boldsymbol{x}\mid\boldsymbol{x}\in\mathcal{D}\}$ for some $0<\epsilon<1$, then the covering number $N(\mathcal{D},\epsilon\mathcal{D})$ obeys [36]

$$\left(\frac{1}{\epsilon}\right)^n\leq N(\mathcal{D},\epsilon\mathcal{D})\leq\left(2+\frac{1}{\epsilon}\right)^n. \tag{21}$$

We can *equivalently* say that the metric entropy $\log N(\mathcal{D},\epsilon\mathcal{D})$ equals $n\log\Theta(1/\epsilon)$. We shall use this notation for specifying metric entropies of key sets in the theorems to follow.

We state a technical lemma needed to prove Theorems 4 and 5. The lemma bounds the difference between norms of topologically close projection operators as a function of the covering resolution, thus providing a characterization of the sets used to cover over the space of interest.

Lemma 5. *Let $\mathcal{G}(m)=\left\{\boldsymbol{Y}\in\mathbb{R}^{m\times 2}\;\middle|\;\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}=\mathbf{I}\right\}$, $\mathcal{D}(m)=\left\{[\boldsymbol{y}_1,\boldsymbol{y}_2]\in\mathbb{R}^{m\times 2}\;\middle|\;\max\limits_{j=1,2}\|\boldsymbol{y}_j\|_2\leq 1\right\}$ and $0<\epsilon<1$. There exists a covering of $\mathcal{G}(m)$ with metric entropy $\leq 2m\log\Theta(1/\epsilon)$ w.r.t. $\epsilon\mathcal{D}(m)$ such that for any $\boldsymbol{Y},\boldsymbol{Z}\in\mathcal{G}(m)$ satisfying $\boldsymbol{Y}-\boldsymbol{Z}\in\epsilon\mathcal{D}(m)$ we have*

$$\left|\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Y})}\boldsymbol{x}\right\|_2-\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z})}\boldsymbol{x}\right\|_2\right|\leq\sqrt{2}\epsilon\|\boldsymbol{x}\|_2 \tag{22}$$

for all $\boldsymbol{x}\in\mathbb{R}^m$.

Proof: Appendix L. ■

Section IV-D provides additional remarks on Lemma 5.

We are now ready to extend Theorem 3 to the case where the complexity of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ is large (possibly infinite). We shall do so for Bernoulli and Gaussian priors (as illustrative distributions) in Theorems 4 and 5 respectively. The proofs for both these theorems follow on the same lines as that of Theorem 3, except that the probability bounds of Lemma 1 are replaced by those of Lemmas 4 and 3 for Bernoulli and Gaussian priors, respectively.

Theorem 4. *Let $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M}=\{\mathbf{0}\}$, the sets $\mathcal{G}(m),\mathcal{G}(n)$ and $\mathcal{D}(m),\mathcal{D}(n)$ be defined according to Lemma 5, and $\boldsymbol{M}=\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\in\mathbb{R}^{m\times n}$ be a rank one random matrix with components of $\boldsymbol{x}$ (respectively $\boldsymbol{y}$) drawn independently from a symmetric Bernoulli distribution with $\mathcal{K}'$ chosen as*

$$\mathcal{K}'=\left\{\lambda\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\;\middle|\;\boldsymbol{x}\in\{-1,1\}^m,\boldsymbol{y}\in\{-1,1\}^n,\lambda\in\mathbb{R}\right\}. \tag{23}$$

and $\mathcal{M}=\mathcal{K}'-\mathcal{K}'$. Let $p_c\log\Theta(1/\epsilon)$ denote the metric entropy of the set $\mathcal{G}(m)\bigcap\{\mathcal{C}(\boldsymbol{X})\mid\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}\}$ w.r.t. $\epsilon\mathcal{D}(m)$, $p_r\log\Theta(1/\epsilon)$ denote the metric entropy of the set $\mathcal{G}(n)\bigcap\{\mathcal{R}(\boldsymbol{X})\mid\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}\}$ w.r.t. $\epsilon\mathcal{D}(n)$, for

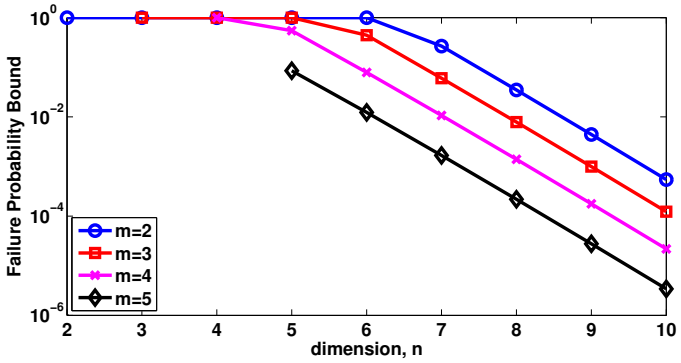

Fig. 2. Exponentially decaying behavior of the theoretically predicted failure probability bound in Theorem 5, *w.r.t.* n for fixed values of m, for parameters $\epsilon = 0.1$, $\delta = 10^{-4}$ and $p = m + n - 3$ for the lifted linear convolution map (p defined as in Theorem 5).

any $1 > \epsilon \geq \epsilon_0 > 0$ *and let* $p = p_c + p_r$. *For any constant* $\delta' \in \left(0, 1 - 2\epsilon^2\right)$, *the sufficient conditions of Theorem 2 are satisfied with probability greater than* $\left(1 - 16\exp\left[p\log\Theta\left(\frac{1}{\epsilon}\right) - (m+n)\frac{1-\delta}{4}\right]\right)$ *with* $\delta = 1 - \left(\sqrt{1-\delta'} - \sqrt{2}\epsilon\right)^2$.

Proof: Appendix M. ∎

Theorem 5. *Let* $\mathcal{N}(\mathscr{S}, 1) = \{\mathbf{0}\}$, *the sets* $\mathcal{G}(m), \mathcal{G}(n)$ *and* $\mathcal{D}(m), \mathcal{D}(n)$ *be defined according to Lemma 5, and* $\boldsymbol{M} = \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \in \mathbb{R}^{m\times n}$ *be a rank one random matrix with components of* $\boldsymbol{x}$ *(respectively* $\boldsymbol{y}$*) drawn independently from a standard Gaussian distribution. Let* $p_c \log\Theta(1/\epsilon)$ *denote the metric entropy of the set* $\mathcal{G}(m)\bigcap\{\mathcal{C}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \setminus \{\mathbf{0}\}\}$ *w.r.t.* $\epsilon\mathcal{D}(m)$, $p_r \log\Theta(1/\epsilon)$ *denote the metric entropy of the set* $\mathcal{G}(n)\bigcap\{\mathcal{R}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \setminus \{\mathbf{0}\}\}$ *w.r.t.* $\epsilon\mathcal{D}(n)$ *for any* $0 < \epsilon < 1$ *and let* $p = p_c + p_r$. *For any constant* $\delta' \in \left(0, 1 - 2\epsilon^2\right)$, *the sufficient conditions of Theorem 2 are satisfied with probability greater than* $\left(1 - C(m,n,\delta)\exp\left[p\log\Theta\left(\frac{1}{\epsilon}\right) - (m+n)\log\frac{1}{\sqrt{\delta}}\right]\right)$ *where*

$$C(m,n,\delta) = \exp\left[2\log mn + 4 - 2\log\frac{2\delta}{1-\delta}\right] = \left(\frac{1}{\delta} - 1\right)^2\Theta\left(m^2n^2\right). \tag{24}$$

and $\delta = 1 - \left(\sqrt{1-\delta'} - \sqrt{2}\epsilon\right)^2$.

Proof: Appendix N. ∎

Sections IV-E and IV-F1 provide additional remarks on Theorems 4 and 5.

A non-trivial illustration of the theoretical scaling law bound of Theorem 5 is provided in Fig. 2, with $\mathscr{S}(\cdot)$ as the lifted linear convolution map. Since the bound is parametrized by (ϵ, δ), we choose $\epsilon = 0.1$ and $\delta = 10^{-4}$ for the illustration. Quite surprisingly (and fortunately), the metric entropy p in Theorem 5 can be exactly characterized when $\mathscr{S}(\cdot)$ represents the lifted linear convolution map. Specifically, we have $p = m + n - 3$. We refer the reader to Proposition 2 in Section V-B for details.

Remark 7. *We can obtain results analogous to Theorems 4 and 5 when* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are drawn from non-identical distributions, e.g.* $\boldsymbol{x}$ *is component-wise i.i.d. symmetric Bernoulli and* $\boldsymbol{y}$ *is component-wise i.i.d. standard Gaussian. The argument is a straightforward modification of the proof.*

IV. Discussion

In this section, we elaborate on the intuitions, ideas, assumptions and subtle implications associated with the main results of this paper that were presented in Section III.

A. A Measure of Geometric Complexity

For the purpose of measuring the size/complexity of $\mathcal{N}(\mathscr{S}, 2)\bigcap\mathcal{M}$ in Theorem 3, we used the cardinality $f_{\mathscr{S},\mathcal{M}}(m,n)$ of the set $\{(\mathcal{C}(\boldsymbol{X}), \mathcal{R}(\boldsymbol{X})) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2)\bigcap\mathcal{M} \setminus \{\mathbf{0}\}\}$ as a surrogate. This set essentially lists the distinct pairs of row and column spaces in the rank two null space of the lifted linear operator $\mathscr{S}(\cdot)$ that are not excluded by the domain restriction $\boldsymbol{M} \in \mathcal{K}'$. We note that the cardinality of the set $\mathcal{N}(\mathscr{S}, 2)\bigcap\mathcal{M}\bigcap\{\boldsymbol{X} \in \mathbb{R}^{m\times n} \mid \|\boldsymbol{X}\|_{\mathrm{F}} = 1\}$ could be infinite while its complexity could be finite in the sense just described. The same measure of complexity is used for the extensions of Theorem 3 in Theorems 4 and 5. Throughout rest of the paper, any reference to the complexity of a set of matrices $\mathcal{M}' \subseteq \mathbb{R}^{m\times n}$ is in the sense just described, *i.e.* through the cardinality of the set $\{(\mathcal{C}(\boldsymbol{X}), \mathcal{R}(\boldsymbol{X})) \mid \boldsymbol{X} \in \mathcal{M}' \setminus \{\mathbf{0}\}\}$.

B. The Role of Conic Prior

There are three distinct aspects to the prior knowledge in terms of the conic constraint $\boldsymbol{M} \in \mathcal{K}'$ on the unknown signal.

1) *Probability Bounds:* A key advantage of prior knowledge about the signal is apparent from the union bounding step in the proof of Theorem 3. Union bounding over the set $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ always gives better bounds than union bounding over the superset $\mathcal{N}(\mathscr{S}, 2) \setminus \{\boldsymbol{0}\}$, the quantitative difference being the number $f_{\mathscr{S},\mathcal{M}}(m, n)$ in the bound of Theorem 3. In general, the difference could be exponentially large in m or n (see Theorems 4 and 5). We also note that $\mathcal{K}'$ does not need to be a cone in order to exploit this approach to improve the probability bounds.
2) *Computational Trade-offs:* Recalling Remark 4, the size of $\mathcal{K}'$ also trades off the ease of computation and the identifiability bounds of Theorem 3. If the size of $\mathcal{K}'$ needs to be increased to ease computation, an effort must be made to not suffer a substantial increase in the size/complexity of the set $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$. For high dimensional problems (m or n is large), non-convex conic priors like the sparse cone in compressed sensing [37] and the low-rank cone in matrix completion [29] have been shown to admit good computationally tractable relaxations.
3) *Geometric Complexity Measure:* The measure of geometric complexity described in Section IV-A followed naturally from Theorem 2 in an effort to describe the identifiability of a BIP in terms of quantities like row and column spaces familiar from linear algebra. This measure of complexity is *invariant w.r.t. conic extensions* in the following way. Let $\mathcal{M}' \subseteq \mathbb{R}^{m \times n}$ be any set of matrices and let $\mathcal{M}''$ denote its conic extension, defined as

$$\mathcal{M}'' \triangleq \left\{ \lambda \boldsymbol{X} \;\middle|\; \boldsymbol{X} \in \mathbb{R}^{m \times n}, \lambda \in \mathbb{R}^{+} \right\}. \tag{25}$$

Then, we have

$$\{(\mathcal{C}(\boldsymbol{X}), \mathcal{R}(\boldsymbol{X})) \mid \boldsymbol{X} \in \mathcal{M}' \setminus \{\boldsymbol{0}\}\} = \{(\mathcal{C}(\boldsymbol{X}), \mathcal{R}(\boldsymbol{X})) \mid \boldsymbol{X} \in \mathcal{M}'' \setminus \{\boldsymbol{0}\}\}. \tag{26}$$

Qualitatively speaking, the flavor of results in this paper could also be derived for non-conic priors but the measure of geometric complexity that is used is implicitly based on conic extensions. Thus, there is no significant loss of generality in restricting ourselves to conic priors.

C. The Role of δ

Although the parameter $\delta \in (0, 1)$ appears in Theorem 3 as an artifact of our proof strategy, it has an important practical consequence. It represents a tolerance parameter for approximate versus exact prior information on the input signals. Specifically, Theorem 3 is a statement about identifiability up to a δ-neighborhood around the true signal $(\boldsymbol{x}, \boldsymbol{y})$. The same holds true for Theorems 4 and 5 describing the large/infinite complexity case.

D. Interpretation of Lemma 5

Lemma 5 can be informally restated as follows. Keeping (22) satisfied, $\mathcal{G}(m)$ can always be covered by $\epsilon\mathcal{D}(m)$ with metric entropy $\leq 2m \log \Theta(1/\epsilon)$. In Theorems 4 and 5 below, we are interested in covering the subset $\mathcal{G}(m) \bigcap \{\mathcal{C}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}\} \subseteq \mathcal{G}(m)$ by $\epsilon\mathcal{D}(m)$ and suppose that the resulting metric entropy is $p_c \log \Theta(1/\epsilon)$. In a sense, Lemma 5 represents the worst case scenario that p_c is upper bounded by $2m$ and no better upper bound is known. In the worst case, the aforementioned subset of $\mathcal{G}(m)$ has nearly the same complexity as $\mathcal{G}(m)$ and this happens when the set $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ does not represent a large enough structural restriction on the set of rank two matrices in $\mathbb{R}^{m \times n}$. For large m, to guarantee identifiability for most inputs, we would (realistically) want p_c to be less than m by at least a constant factor. This is implied by Theorems 4 and 5. Informally, smaller or more structured $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ implies a smaller value of p_c which in turn implies identifiability for a greater fraction of the input ensemble.

E. The Gaussian and Bernoulli Special Cases

A standard Gaussian prior on the elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ gives an example of the set $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ with infinite complexity, provided that $\mathcal{N}(\mathscr{S}, 2)$ is complex enough. In this case, $\mathcal{K} = \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n\}$ in Problem (P_2) implying that $\mathcal{K}' \supseteq \{\boldsymbol{W} \in \mathbb{R}^{m \times n} \mid \operatorname{rank}(\boldsymbol{W}) \leq 1\}$ from (8). Thus, $\mathcal{M} \supseteq \{\boldsymbol{W} \in \mathbb{R}^{m \times n} \mid \operatorname{rank}(\boldsymbol{W}) \leq 2\} \supseteq \mathcal{N}(\mathscr{S}, 2)$ and hence $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \mathcal{N}(\mathscr{S}, 2)$. Since $\mathcal{M}$ is superfluous in this case, Theorem 5 omits all references to it. If the row or column spaces of matrices in $\mathcal{N}(\mathscr{S}, 2)$ are parametrized by one or more real parameters (see Section V-B for an example involving the linear convolution operator), then $\mathcal{N}(\mathscr{S}, 2)$ has infinite complexity.

The scenario of a Bernoulli prior on elements of $\boldsymbol{x}$ and $\boldsymbol{y}$ gives an example of the set $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ with finite (but exponentially large in $m+n$) complexity, provided that $\mathcal{N}(\mathscr{S}, 2)$ is complex enough. The precise statement requires a little more care than the Gaussian case described above. The motivation behind considering Bernoulli priors is to restrict the unit vectors $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ and $\boldsymbol{y}/\|\boldsymbol{y}\|_2$ to take values from a large but finite set while adhering to the requirement of a conic prior on $(\boldsymbol{x}, \boldsymbol{y})$ according to Problem (P_2). Thus, in this case we have $\mathcal{K} = \{(\lambda_1 \boldsymbol{x}, \lambda_2 \boldsymbol{y}) \mid \boldsymbol{x} \in \{-1, 1\}^m, \boldsymbol{y} \in \{-1, 1\}^n, \lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{R}\}$. Let us select $\mathcal{K}'$ according to (8), but without any relaxation, as

$$\mathcal{K}' = \left\{ \lambda \boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\middle|\; \boldsymbol{x} \in \{-1, 1\}^m, \boldsymbol{y} \in \{-1, 1\}^n, \lambda \in \mathbb{R} \right\}. \tag{27}$$

Clearly, matrices in $\mathcal{K}'$ can account for at most 2^{m-1} distinct column spaces and 2^{n-1} distinct row spaces, thus implying that matrices in $\mathcal{M} = \mathcal{K}' - \mathcal{K}'$ are generated by at most $\binom{2^{m-1}}{2} \leq 2^{2m-2}$ distinct column spaces and at most $\binom{2^{n-1}}{2} \leq 2^{2n-2}$ distinct row spaces. Thus, $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \subseteq \mathcal{M}$ is of finite complexity. It is clear that the complexity of $\mathcal{M}$ is $\exp(\Theta(m+n))$ so that if $\mathcal{M} \setminus \mathcal{N}(\mathscr{S}, 2)$ is small enough then the complexity of $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ is exponentially large in $m+n$.

F. Distinctions between Theorems 4 and 5

1) Assumptions on ϵ: We prevent an arbitrarily small ϵ for Theorem 4 by imposing a strictly positive lower bound $\epsilon_0 > 0$. This is necessary for Bernoulli priors on $\boldsymbol{x}$ and $\boldsymbol{y}$ since $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$ has a finite complexity, implying that the covering numbers of $\mathcal{G}(m) \bigcap \{\mathcal{C}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \setminus \{\boldsymbol{0}\}\}$ *w.r.t.* $\epsilon\mathcal{D}(m)$ and $\mathcal{G}(n) \bigcap \{\mathcal{R}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \setminus \{\boldsymbol{0}\}\}$ *w.r.t.* $\epsilon\mathcal{D}(n)$ have an absolute upper bound independent of ϵ. Thus, the logarithmic dependence (of the key metric entropies) on $1/\epsilon$ cannot hold unless ϵ is lower bounded away from zero. Theorem 5, in contrast, allows for arbitrarily small ϵ since $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \mathcal{N}(\mathscr{S}, 2)$ has infinite complexity, for Gaussian priors on $\boldsymbol{x}$ and $\boldsymbol{y}$. Despite this distinction between Theorems 4 and 5, we choose to present our results in the stated form to emphasize similarity in the theorem statements and proofs.

2) A constant factor loss: We loose a constant factor of approximately 2 in the exponent on the r.h.s. of (20) as compared to (19) for a fixed $\delta \in (0, 1)$ (compared using first order approximation of $\log \delta$). While this seems to be an artifact of the proof strategy, it is unclear whether a better constant can be obtained for the symmetric Bernoulli distribution (or more generally, for subgaussian distributions [38]). Indeed, for the proof of Lemma 3 in Appendix J, we have used the rotational invariance property of the multivariate standard normal distribution. This property does not carry over to general subgaussian distributions.

V. Numerical Results on Blind Deconvolution

We observe that if $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$ then $f_{\mathscr{S},\mathcal{M}}(m, n) = 0$ and Theorem 3 correctly predicts that the input signals are identifiable with probability one (in agreement with Proposition 1). Below, we consider example bilinear maps and input distributions with $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \neq \{\boldsymbol{0}\}$ and numerically examine the scaling behavior suggested by Lemmas 1, 4 and 3 and Theorems 3, 4 and 5. Since Lemma 1 and Theorem 3 impose only broad constraints on the input distribution, for the purpose of numerical simulations, we construct a specific input distribution that satisfies assumptions (A1)-(A3) in Section V-A. Since this research was motivated by our interest to understand the cone constrained blind deconvolution problem (P_1), our selection of example bilinear maps are closely related to the linear convolution map. We provide a partial description of the rank two null space for the linear convolution map in Section V-B.

A. Bi-orthogonally Supported Uniform Distributions

A bi-orthogonal set of vectors is a collection of orthonormal vectors and their additive inverses. It is widely used for signal representation in image processing and as a modulation scheme in communication systems. We can construct a uniform distribution over a bi-orthogonal set and it would satisfy assumptions (A1)-(A3) as shown below.

Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m\}$ be an orthonormal basis for $\mathbb{R}^m$ and the random unit vector $\boldsymbol{u} \in \{\pm\boldsymbol{e}_1, \pm\boldsymbol{e}_2, \ldots, \pm\boldsymbol{e}_m\}$ be drawn according to the law

$$\Pr(\boldsymbol{u} = +\boldsymbol{e}_j) = \Pr(\boldsymbol{u} = -\boldsymbol{e}_j) = \frac{1}{2m}, \quad \forall 1 \leq j \leq m, \tag{28}$$

where $\boldsymbol{u}$ has the same meaning as in Lemma 1 and Theorem 3. Let $\|\boldsymbol{x}\|_2$ be drawn from a distribution (independent of $\boldsymbol{u}$) supported on the non-negative real axis with $\mathrm{E}\left[\|\boldsymbol{x}\|_2^2\right] = m$. Then, by construction, $\boldsymbol{x} = \|\boldsymbol{x}\|_2 \cdot \boldsymbol{u}$ satisfies assumption (A3) and it also satisfies assumption (A2) if $\|\boldsymbol{y}\|_2$ and $\boldsymbol{v}$ are drawn analogously but independent of $\boldsymbol{u}$ and $\|\boldsymbol{x}\|_2$. Using (28), we further observe that

$$\mathrm{E}[\boldsymbol{u}] = \sum_{j=1}^{m} [\Pr(\boldsymbol{u} = +\boldsymbol{e}_j) - \Pr(\boldsymbol{u} = -\boldsymbol{e}_j)] \cdot \boldsymbol{e}_j = \boldsymbol{0} \tag{29}$$

and,

$$\mathrm{E}\left[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right] = \sum_{j=1}^{m} [\Pr(\boldsymbol{u} = +\boldsymbol{e}_j) + \Pr(\boldsymbol{u} = -\boldsymbol{e}_j)] \cdot \boldsymbol{e}_j\boldsymbol{e}_j^{\mathrm{T}} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{e}_j\boldsymbol{e}_j^{\mathrm{T}} = \frac{1}{m}\mathbf{I} \tag{30}$$

where the last equality in (30) is true since $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m\}$ is an orthonormal basis for $\mathbb{R}^m$. By independence of $\|\boldsymbol{x}\|_2$ from $\boldsymbol{u}$ we have

$$\mathrm{E}[\boldsymbol{x}] = \mathrm{E}[\|\boldsymbol{x}\|_2] \cdot \mathrm{E}[\boldsymbol{u}] = \boldsymbol{0} \tag{31}$$

from (29), and

$$\mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] = \mathrm{E}\left[\|\boldsymbol{x}\|_2^2\right] \cdot \mathrm{E}\left[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\right] = m \cdot \frac{1}{m}\mathbf{I} = \mathbf{I} \tag{32}$$

from (30). Hence, $\boldsymbol{x}$ is a zero mean random vector with an identity covariance matrix and thus satisfies assumption (A1).

Following the same line of reasoning as in Section IV-E, we can show that a bi-orthogonally supported uniform prior on $\boldsymbol{x}$ and $\boldsymbol{y}$ gives an example of the set $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ with small complexity in the sense described in Section IV-A. Indeed, we have $\mathcal{K}=\{(\lambda_1\boldsymbol{e}_i,\lambda_2\boldsymbol{f}_j)\mid 1\leq i\leq m,1\leq j\leq n,\lambda_1\in\mathbb{R},\lambda_2\in\mathbb{R}\}$ where $\{\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots,\boldsymbol{e}_m\}$ and $\{\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_n\}$ respectively form an orthonormal basis for $\mathbb{R}^m$ and $\mathbb{R}^n$. Let us select $\mathcal{K}'$ according to (8), but without any relaxation, as $\mathcal{K}'=\left\{\lambda\boldsymbol{e}_i\boldsymbol{f}_j^{\mathrm{T}}\;\middle|\;1\leq i\leq m,1\leq j\leq n,\lambda\in\mathbb{R}\right\}$. It is clear that matrices in $\mathcal{K}'$ can account for at most m distinct column spaces and n distinct row spaces, thus implying that matrices in $\mathcal{M}=\mathcal{K}'-\mathcal{K}'$ are generated by at most $\binom{m}{2}\leq m^2$ distinct column spaces and by at most $\binom{n}{2}\leq n^2$ distinct row spaces. Thus, $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\subseteq\mathcal{M}$ is of small complexity (only polynomially large in m and n). In fact, *exhaustive search for Problem* (P$_4$) *is tractable for any bi-orthogonally supported uniform prior, owing to the small complexity of* $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$.

B. Null Space of Linear Convolution

The following proposition establishes a parametric representation of a subset of $\mathcal{N}(\mathscr{S},2)$ where $\mathscr{S}\colon\mathbb{R}^{m\times n}\to\mathbb{R}^{m+n-1}$ denotes the lifted equivalent of the linear convolution map in Problem (P$_1$). As described by (10) in Section II-C, let $\boldsymbol{S}_k\in\mathbb{R}^{m\times n}$, $1\leq k\leq m+n-1$ denote a basis for $\mathscr{S}(\cdot)$. For $1\leq i\leq m$, $1\leq j\leq n$ and $1\leq k\leq m+n-1$, we have the description

$$(\boldsymbol{S}_k)_{ij}=\begin{cases}1, & i+j=k+1,\\ 0, & \text{otherwise.}\end{cases}\tag{33}$$

Fig. 1 illustrates a toy example of the linear convolution map with $m=3$ and $n=4$.

Proposition 2. *If* $\boldsymbol{X}\in\mathbb{R}^{m\times n}$ *admits a factorization of the form*

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{u} & 0\\ 0 & -\boldsymbol{u}\end{bmatrix}\begin{bmatrix}0 & \boldsymbol{v}^{\mathrm{T}}\\ \boldsymbol{v}^{\mathrm{T}} & 0\end{bmatrix}\tag{34}$$

for some $\boldsymbol{v}\in\mathbb{R}^{n-1}$ *and* $\boldsymbol{u}\in\mathbb{R}^{m-1}$, *then* $\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)$.

Proof: Appendix O. ∎

Since the set of $m\times n$ dimensional rank two matrices has $2(m+n-2)$ DoF and $\mathscr{S}(\cdot)$ maps $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m+n-1}$ with $\mathcal{N}(\mathscr{S},1)=\{\boldsymbol{0}\}$, $\mathcal{N}(\mathscr{S},2)$ has at most $(2m+2n-4)-(m+n-1)=(m+n-3)$ DoF. We see that the representation on the r.h.s. of (34) also has $(m+n-3)$ DoF, so that our parametrization is tight up to DoF. The converse of Proposition 2 is false in general [11].

C. Verification Methodology

We test identifiability by (approximately) solving the following optimization problem,

$$\begin{aligned}\underset{\boldsymbol{X}}{\text{minimize}}\quad & \operatorname{rank}(\boldsymbol{X})\\ \text{subject to}\quad & \|\boldsymbol{X}-\boldsymbol{M}\|_{\mathrm{F}}\leq\mu,\\ & \mathscr{S}(\boldsymbol{X})=\boldsymbol{0},\end{aligned}\tag{P$_5$}$$

where $\boldsymbol{M}=\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}$ is the true matrix and ϵ is a tuning parameter. The rationale behind solving Problem (P$_5$) is as follows. If the sufficient conditions of Theorem 2 are not satisfied, then $\exists\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ such that both $\boldsymbol{x}\in\mathcal{C}(\boldsymbol{X})$ and $\boldsymbol{y}\in\mathcal{R}(\boldsymbol{X})$ are true. We approximate the event

$$\mathcal{E}_1=\{\exists\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\text{ with }\boldsymbol{u}\in\mathcal{C}(\boldsymbol{X}),\boldsymbol{v}\in\mathcal{R}(\boldsymbol{X})\}\tag{35a}$$

by the event

$$\mathcal{E}_2=\{\exists\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\text{ such that }\|\boldsymbol{X}-\boldsymbol{M}\|_{\mathrm{F}}\leq\mu\}.\tag{35b}$$

As Problem (P$_5$) is itself NP-hard to solve exactly, we can employ the re-weighted nuclear norm heuristic [39] to solve Problem (P$_5$) approximately. If the resulting solution to Problem (P$_5$) has rank two then we declare that event $\mathcal{E}_2$ has happened. Clearly, we have $\mathcal{E}_2\subseteq\mathcal{E}_1$ so that sufficient conditions for identifiability by Theorem 2 fail if event $\mathcal{E}_2$ took place.

The examples we consider in Sections V-D to V-F are, however, motivated from the representation in (34) and share the same parametrization structure for $\mathcal{N}(\mathscr{S},2)$. This enables us to use approximate verification techniques that are faster than the re-weighted nuclear norm heuristic, especially if the search space is discrete and finite. The re-weighted nuclear norm heuristic is still useful if no parametrization structure is available for $\mathcal{N}(\mathscr{S},2)$.

D. Small Complexity of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$

Let $\boldsymbol{x}\in\{\boldsymbol{e}'_1,\boldsymbol{e}'_2,\ldots,\boldsymbol{e}'_m\}$ and $\boldsymbol{y}\in\{\boldsymbol{f}'_1,\boldsymbol{f}'_2,\ldots,\boldsymbol{f}'_n\}$ be drawn from bi-orthogonally supported uniform distributions, as described in Section V-A, where $\{\boldsymbol{e}'_1,\boldsymbol{e}'_2,\ldots,\boldsymbol{e}'_m\}$ and $\{\boldsymbol{f}'_1,\boldsymbol{f}'_2,\ldots,\boldsymbol{f}'_n\}$ respectively represent the canonical bases for $\mathbb{R}^m$ and

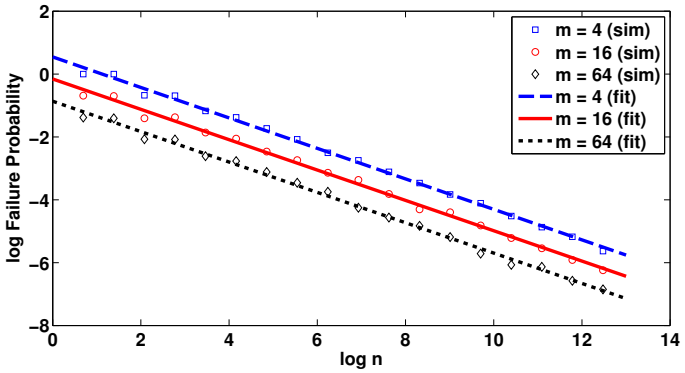

Fig. 3. Linear Scaling behavior of $\log$(Failure Probability) with $\log n$ for fixed values of m. The absolute value of the fitted slope is 0.48.

$\mathbb{R}^n$. We consider a lifted linear operator $\mathscr{S}(\cdot)$ with the following description: $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ consists of $\lfloor\sqrt{m}\rfloor\cdot\lfloor\sqrt{n}\rfloor$ parts and the $(i,j)^{\text{th}}$ part $\mathcal{P}_{ij}$, $1\le i\le\lfloor\sqrt{m}\rfloor$, $1\le j\le\lfloor\sqrt{n}\rfloor$ is given by

$$\mathcal{P}_{ij}=\left\{\lambda\begin{bmatrix}\boldsymbol{e}_i & 0\\ 0 & -\boldsymbol{e}_i\end{bmatrix}\begin{bmatrix}0 & \boldsymbol{f}_j^{\mathrm{T}}\\ \boldsymbol{f}_j^{\mathrm{T}} & 0\end{bmatrix}\;\middle|\;\lambda\in\mathbb{R}\right\} \tag{36}$$

where $\{\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots,\boldsymbol{e}_{m-1}\}$ and $\{\boldsymbol{f}_1,\boldsymbol{f}_2,\ldots,\boldsymbol{f}_{n-1}\}$ respectively denote the canonical basis for $\mathbb{R}^{m-1}$ and $\mathbb{R}^{n-1}$, and $\lfloor\cdot\rfloor$ is the floor function. Clearly, the elements of $\mathcal{P}_{ij}$ are closely related to the representation in (34). For this lifted linear operator, the bound of Theorem 3 is applicable with $f_{\mathscr{S},\mathcal{M}}(m,n)=\lfloor\sqrt{m}\rfloor\cdot\lfloor\sqrt{n}\rfloor$ implying that the probability of failure to satisfy the sufficient conditions of Theorem 2 decreases as $O(1/\sqrt{mn})$. Since exhaustive search for event $\mathcal{E}_2$ is tractable (see Section V-A), we employ the same to compute the failure probability. The results are plotted in Fig. 3 on a log-log scale. Note that we have plotted the best linear fit for the simulated parameter values, since the probabilities can be locally discontinuous in $\log n$ due to the appearance of $\lfloor\cdot\rfloor$ function in the expression of $f_{\mathscr{S},\mathcal{M}}(m,n)$. We see that the simulated order of growth of the failure probability is $O\big(n^{-0.48}\big)$ for every fixed value of m (exponent determined by slope of plot in Fig. 3) almost exactly matches the theoretically predicted order of growth (equals $O\big(n^{-0.5}\big)$).

E. Large Complexity of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$

Let $\boldsymbol{x}\in\mathbb{R}^m$ and $\boldsymbol{y}\in\mathbb{R}^n$ be drawn component-wise independently from a symmetric Bernoulli distribution (see Section IV-E) and let $\tau\in(0,1)$ be a constant. Following our guiding representation (34), we consider a lifted linear operator $\mathscr{S}(\cdot)$ with the following description: $\mathcal{N}(\mathscr{S},2)$ consists of $2^{\lfloor\tau m\rfloor}\times 2^{\lfloor\tau n\rfloor}$ parts and the $(i,j)^{\text{th}}$ part $\mathcal{P}_{ij}$, $0\le i\le 2^{\lfloor\tau m\rfloor}-1$, $0\le j\le 2^{\lfloor\tau n\rfloor}-1$, is given by

$$\mathcal{P}_{ij}=\left\{\lambda\begin{bmatrix}\boldsymbol{g}_i & 0\\ \mathbf{1} & -\boldsymbol{g}_i\\ 0 & -\mathbf{1}\end{bmatrix}\begin{bmatrix}0 & \boldsymbol{h}_j^{\mathrm{T}} & \mathbf{1}^{\mathrm{T}}\\ \boldsymbol{h}_j^{\mathrm{T}} & \mathbf{1}^{\mathrm{T}} & 0\end{bmatrix}\;\middle|\;\lambda\in\mathbb{R}\right\} \tag{37}$$

where $\boldsymbol{g}_i\in\{-1,1\}^{\lfloor\tau m\rfloor}$ (respectively $\boldsymbol{h}_j\in\{-1,1\}^{\lfloor\tau n\rfloor}$) denotes the binary representation of i (respectively j) of length $\lfloor\tau m\rfloor$ bits (respectively $\lfloor\tau n\rfloor$ bits) expressed in the alphabet set $\{-1,1\}$, and the all one column vectors in (37) are of appropriate dimensions so that the elements of $\mathcal{P}_{ij}$ are matrices in $\mathbb{R}^{m\times n}$. The bound in Theorem 4 is applicable to this example. We employ exhaustive search for event $\mathcal{E}_2$ for small values of m and n (it is computationally intractable for large m or n). The results are plotted in Fig. 4 on a semilog scale, where we have used $\tau=0.2$ and $\delta'=0.3$ and δ' is as in the statement of Theorem 4. As in the case of Fig. 3, we plot the best linear fit for the simulated parameter values to disregard local discontinuities introduced due to the use of the $\lfloor\cdot\rfloor$ function.

Since it is hard to analytically compute the metric entropies p_c and p_r, we shall settle for a numerical verification of the scaling law with problem dimension and an approximate argument as to the validity of predictions made by Theorem 4 for this example. By construction, we have the bounds $p_c\le\lfloor\tau m\rfloor$ and $p_r\le\lfloor\tau n\rfloor$ but the careful reader will note that because of the element-wise constant magnitude property of a symmetric Bernoulli random vector, it does not lie in the column span of any of the matrices in $\mathcal{N}(\mathscr{S},2)$, as described by the generative description in (37), but can be arbitrarily close to such a span as m increases. We thus expect that $p_c=\epsilon_c m$ and $p_r=\epsilon_r n$ for some parameters ϵ_c and ϵ_r close to zero. By choice of parameters, $\epsilon\le\sqrt{(1-\delta')/2}=0.59$. With $\epsilon_c=\epsilon_r=0$ and setting $\epsilon=0.01$ the theoretical prediction on the absolute value of the slope is 0.073 which is quite close to the simulated value of 0.093. We clearly recover the linear scaling behavior of the logarithm of failure probability with the problem dimension n.

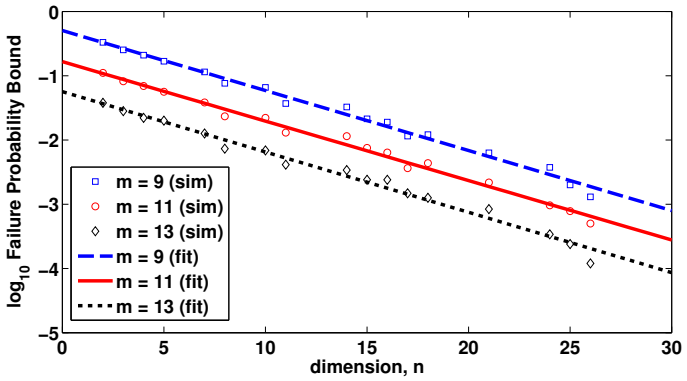

Fig. 4. Linear Scaling behavior of log(Failure Probability) with problem dimension n for fixed values of m. The absolute value of the fitted slopes are between 0.093 and 0.094.

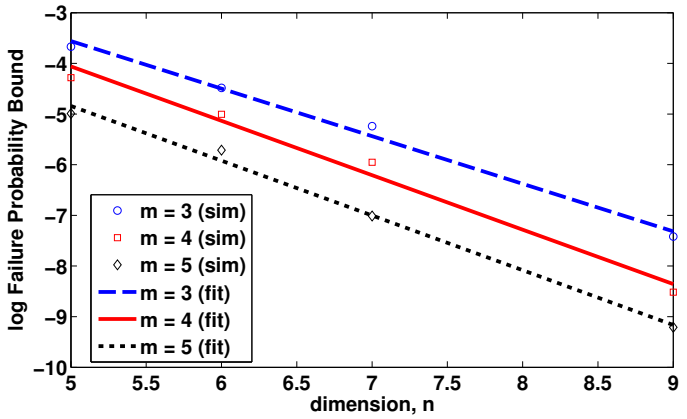

Fig. 5. Exponentially decaying behavior of the simulated failure probability *w.r.t.* n for fixed values of m, for parameter $\mu = 0.8$ and the lifted linear convolution map. The absolute value of the fitted slopes are between 0.94 and 1.08.

F. Infinite Complexity of $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M}$

Let $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$ be drawn component-wise independently from a standard Normal distribution. We consider the linear convolution operator from Problem (P_1), letting $\mathscr{S}(\cdot)$ denote the lifted linear convolution map. A representation of $\mathscr{S}(\cdot)$ and a description of the rank two null space $\mathcal{N}(\mathscr{S}, 2)$ has been mentioned in the prequel (Section V-B). The bound in Theorem 5 is applicable to this example. However, unlike the examples in Sections V-D and V-E, we cannot employ exhaustive search over $\mathcal{N}(\mathscr{S}, 2)$ to test identifiability, since the search space is uncountably infinite by Proposition 2. We resort to the method described in Section V-C relying on the re-weighted nuclear norm heuristic. The results are plotted in Fig. 5 on a semilog scale, where we have used $\mu = 0.8$ to detect the occurrence of the event $\mathcal{E}_2$ as described by (35b), and $\boldsymbol{M}$ in Problem (P_5) is normalized such that $\|\boldsymbol{M}\|_{\mathrm{F}} = 1$. A relatively high value of $\mu = 0.8$ is used to ensure that the rare event $\mathcal{E}_2$ admits a large enough probability of occurrence. Only data points that satisfy $n \geq m$ are plotted since the behavior of the convolution operator is symmetric *w.r.t.* the order of its inputs. Since the re-weighted nuclear norm heuristic does not always converge monotonically in a small number of steps, we stopped execution after a finite number of steps, which might explain the small deviation from linearity, observed in Fig. 5, as compared to the respective best linear fits on the same plot. Nonetheless, we approximately recover the theoretically predicted qualitative linear scaling law of the logarithm of the failure probability with the problem dimension n, for fixed values of m. There does not seem to be an easy way of comparing the constants involved in the simulated result to their theoretical counterparts as predicted by Theorem 5.

VI. Conclusions

Bilinear transformations occur in a number of signal processing problems like linear and circular convolution, matrix product, linear mixing of multiple sources, *etc.* Identifiability and signal reconstruction for the corresponding inverse problems are important in practice and identifiability is a precursor to establishing any form of reconstruction guarantee. In the current work, we determined a series of sufficient conditions for identifiability in conic prior constrained Bilinear Inverse Problems (BIPs) and investigated the probability of achieving those conditions under three classes of random input signal ensembles, *viz.* dependent but uncorrelated, independent Gaussian, and independent Bernoulli. The theory is *unified* in the sense that it is applicable to all BIPs, and is specifically developed for bilinear maps over vector pairs with non-trivial rank two null space. Universal

identifiability is absent for many interesting and important BIPs owing to the non-triviality of the rank two null space, but a deterministic characterization of the input instance identifiability is still possible (may be hard to check). Our probabilistic results were formulated as scaling laws that trade-off probability of identifiability with the complexity of the restricted rank two null space of the bilinear map in question, and results were derived for three different levels of complexity, *viz.* small (polynomial in the signal dimension), large (exponential in the signal dimension) and infinite. In each case, identifiability can hold with high probability depending on the relative geometry of the null space of the bilinear map and the signal space. Overall, most random input instances are identifiable, with the probability of identifiability scaling inversely with the complexity of the rank two null space of the bilinear map. An especially appealing aspect of our approach is that the rank two null space can be partly or fully characterized for many bilinear problems of interest. We demonstrated this by partly characterizing the rank two null space of the linear convolution map, and presented numerical verification of the derived scaling laws on examples that were based on variations of the blind deconvolution problem, exploiting the representation of its rank two null space. Overall, the results in this paper indicate that lifting is a powerful technique for identifiability analysis of general cone constrained BIPs.

REFERENCES

[1] S. Choudhary and U. Mitra, "On Identifiability in Bilinear Inverse Problems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 4325–4329.

[2] ——, "Identifiability Bounds for Bilinear Inverse Problems," in *47th Asilomar Conference on Signals, Systems and Computers*, Nov. 2013, pp. 1677–1681.

[3] J. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 476–488, 2003.

[4] P. D. O'grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.

[5] Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin, "Dictionary learning for noisy and incomplete hyperspectral images," *SIAM J. Imaging Sci.*, vol. 5, no. 1, pp. 33–56, 2012.

[6] D. L. Donoho and V. Stodden, "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?" in *NIPS*, 2003.

[7] C. R. Johnson, Jr., P. Schniter, T. J. Endres, J. D. Behm, D. R. Brown, and R. A. Casas, "Blind Equalization Using the Constant Modulus Criterion: A Review," *Proc. IEEE*, vol. 86, no. 10, pp. 1927–1950, Oct. 1998.

[8] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The Convex Geometry of Linear Inverse Problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[9] S. Choudhary and U. Mitra, "Sparse recovery from convolved output in underwater acoustic relay networks," in *2012 Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2012, pp. 1–8. [Online]. Available: http://www.apsipa.org/proceedings_2012/papers/349.pdf

[10] ——, "Sparse Blind Deconvolution: What Cannot Be Done," in *2014 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2014, pp. 3002–3006.

[11] ——, "On Identifiability Limits for Sparse Blind Deconvolution," 2014, in preparation. [Online]. Available: http://www-scf.usc.edu/~sunavcho/SBD_limit.pdf

[12] E. Balas, "Projection, lifting and extended formulation in integer and combinatorial optimization," *Ann. Oper. Res.*, vol. 140, pp. 125–161, 2005.

[13] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming," *Comm. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013. [Online]. Available: http://dx.doi.org/10.1002/cpa.21432

[14] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inform. Theory*, vol. 60, no. 3, pp. 1711–1732, 2014.

[15] M. S. Asif, W. Mantzel, and J. K. Romberg, "Random Channel Coding and Blind Deconvolution," in *47th Annual Allerton Conference on Communication, Control, and Computing, 2009. Allerton 2009.*, Oct. 2009, pp. 1021–1025.

[16] C. Hegde and R. G. Baraniuk, "Sampling and Recovery of Pulse Streams," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1505–1517, 2011.

[17] P. Walk and P. Jung, "Compressed Sensing on the Image of Bilinear Maps," in *IEEE International Symposium on Information Theory Proceedings (ISIT), 2012*, Jul. 2012, pp. 1291–1295.

[18] D. Gross, "Recovering Low-Rank Matrices From Few Coefficients in Any Basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[19] B. Recht, W. Xu, and B. Hassibi, "Null space conditions and thresholds for rank minimization," *Math. Program.*, vol. 127, no. 1, Ser. B, pp. 175–202, 2011.

[20] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[21] K. Lee, Y. Wu, and Y. Bresler, "Near Optimal Compressed Sensing of Sparse Rank-One Matrices via Sparse Power Factorization," *CoRR*, vol. abs/1312.0525, 2013.

[22] K. Jaganathan, S. Oymak, and B. Hassibi, "Sparse Phase Retrieval: Convex Algorithms and Limitations," *CoRR*, vol. abs/1303.4128, 2013.

[23] ——, "Sparse Phase Retrieval: Uniqueness Guarantees and Recovery Algorithms," *CoRR*, vol. abs/1311.2745, 2013.

[24] A. Beck, "Convexity properties associated with nonconvex quadratic matrix functions and applications to quadratic programming," *J. Optim. Theory Appl.*, vol. 142, no. 1, pp. 1–29, 2009.

[25] A. Kammoun, A. Aissa El Bey, K. Abed-Meraim, and S. Affes, "Robustness of blind subspace based techniques using ℓ_p quasi-norms," in *Signal Processing Advances in Wireless Communications (SPAWC), 2010 IEEE Eleventh International Workshop on*, 2010, pp. 1–5.

[26] R. Gribonval and K. Schnass, "Dictionary identification—sparse matrix-factorization via ℓ_1-minimization," *IEEE Trans. Inform. Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[27] A. Agarwal, A. Anandkumar, and P. Netrapalli, "Exact Recovery of Sparsely Used Overcomplete Dictionaries," *ArXiv e-prints*, vol. abs/1309.1952, Sep. 2013. [Online]. Available: http://arxiv.org/abs/1309.1952

[28] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind System Identification," *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug. 1997.

[29] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[30] E. J. Candès and T. C. Tao, "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[31] F. Kiraly and R. Tomioka, "A Combinatorial Algebraic Approach for the Identifiability of Low-Rank Matrix Completion," *ArXiv e-prints*, vol. abs/1206.6470, Jun. 2012. [Online]. Available: http://arxiv.org/abs/1206.6470

[32] W. Rudin, *Real and complex analysis*, 3rd ed. New York: McGraw-Hill Book Co., 1987.

[33] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[34] M. Ledoux and M. Talagrand, *Probability in Banach Spaces*, ser. Classics in Mathematics. Berlin: Springer-Verlag, 2011, isoperimetry and Processes, Reprint of the 1991 Edition.

[35] E. J. Candès and T. C. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.
[36] R. Vershynin. (2009) Lectures in Geometric Functional Analysis. [Online]. Available: http://www-personal.umich.edu/~romanv/papers/GFA-book/GFA-book.pdf
[37] D. L. Donoho, "Compressed Sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[38] R. Baraniuk, M. A. Davenport, M. F. Duarte, and C. Hegde. (2011, Apr.) An Introduction to Compressive Sensing. Connexions. [Online]. Available: http://cnx.org/content/col11133/1.5/
[39] K. Mohan and M. Fazel, "Reweighted nuclear norm minimization with application to system identification," in *American Control Conference (ACC), 2010*, 2010, pp. 2953–2959.

Appendix A
Proof of Theorem 1

1) Suppose that $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{K}$ is a solution to Problem (P_2) for a given observation $\boldsymbol{z} = \boldsymbol{z}_0 = \boldsymbol{S}(\boldsymbol{x}_0, \boldsymbol{y}_0)$. Setting $\boldsymbol{W}_0 = \boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}$ and using (8) we have $\boldsymbol{W}_0 \in \mathcal{K}'$. Using (11) we get $\mathscr{S}(\boldsymbol{W}_0) = \boldsymbol{S}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \boldsymbol{z}_0$ and $\mathrm{rank}(\boldsymbol{W}_0) = \mathrm{rank}\big(\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}\big) \leq 1$. Thus, $\boldsymbol{W}_0$ is a feasible point of Problem (P_4) with rank at most one. As there exists a rank ≤ 1 matrix $\boldsymbol{W}$ satisfying $\mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}$ and $\boldsymbol{W} \in \mathcal{K}'$, the solution of Problem (P_4) must be of rank one or less.
2) Any $\boldsymbol{W} \in \mathcal{K}'_{\text{opt}} \subseteq \mathcal{K}'$ satisfies $\mathrm{rank}(\boldsymbol{W}) \leq 1$ (see proof of first part) and $\mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}$. Thus,

$$\begin{aligned}
\mathcal{K}'_{\text{opt}} &\subseteq \mathcal{K}' \bigcap \big\{\boldsymbol{W} \in \mathbb{R}^{m\times n} \;\big|\; \mathrm{rank}(\boldsymbol{W}) \leq 1\big\} \bigcap \big\{\boldsymbol{W} \in \mathbb{R}^{m\times n} \;\big|\; \mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}\big\} && \text{(38a)}\\
&= \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\big\} \bigcap \{\boldsymbol{W} \mid \mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}\} && \text{(38b)}\\
&= \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\big\} \bigcap \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; \mathscr{S}\big(\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\big) = \boldsymbol{z}\big\}\\
&= \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\big\} \bigcap \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; \boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{z}\big\} && \text{(38c)}\\
&= \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}, && \text{(38d)}
\end{aligned}$$

where (38b) is due to (8), (38c) is due to (11) and (38d) is true because Problem (P_2) is a feasibility problem.
3) The feasible set for Problem (P_4) is $\mathcal{K}' \bigcap \{\boldsymbol{W} \in \mathbb{R}^{m\times n} \mid \mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}\}$. From the proof of second part, we know that

$$\mathcal{K}' \bigcap \{\boldsymbol{W} \mid \mathrm{rank}(\boldsymbol{W}) \leq 1\} \bigcap \{\boldsymbol{W} \mid \mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}\} = \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}. \tag{39}$$

Thus, clearly

$$\big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\} \subseteq \mathcal{K}' \bigcap \{\boldsymbol{W} \mid \mathscr{S}(\boldsymbol{W}) = \boldsymbol{z}\}. \tag{40}$$

We shall prove the contrapositive statement in each direction. First assume that $\{\boldsymbol{0}\} \subsetneq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}$. By (40), $\boldsymbol{0}$ is a feasible point for Problem (P_4) and thus $\mathrm{rank}(\boldsymbol{0}) = 0$ is the optimal value for this problem. Since every $\boldsymbol{W} \in \mathbb{R}^{m\times n} \setminus \{\boldsymbol{0}\}$ has a rank strictly greater than zero we conclude that $\mathcal{K}'_{\text{opt}} = \{\boldsymbol{0}\} \neq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}$. Conversely, suppose that $\mathcal{K}'_{\text{opt}} \neq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}$. Since $\mathcal{K}'_{\text{opt}} \subseteq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}$ (see proof of second part), $\exists (\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathcal{K}_{\text{opt}}$ such that $\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}} \notin \mathcal{K}'_{\text{opt}}$. By (40), $\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}$ is a feasible point for Problem (P_4) and hence the optimal value of this problem is strictly less than $\mathrm{rank}\big(\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}\big) \leq 1$. The only way for this to be possible is to have $\mathrm{rank}\big(\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}\big) = 1$ and the optimal value of Problem (P_4) as zero. Since the only matrix of rank zero is the all zero matrix, we conclude that $\{\boldsymbol{0}\} = \mathcal{K}'_{\text{opt}} \subsetneq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}\big\}$.

Appendix B
Proof of Corollary 1

From (3), (38a) and (38d) we have

$$\mathcal{K}' \bigcap \mathcal{N}(\mathscr{S}, 1) = \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}(\boldsymbol{0})\big\}. \tag{41}$$

We shall prove the contrapositive statements. First assume that $\{\boldsymbol{0}\} \subsetneq \mathcal{K}' \bigcap \mathcal{N}(\mathscr{S}, 1)$. Using (41), we have $\{\boldsymbol{0}\} \subsetneq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}(\boldsymbol{0})\big\}$ and the last part of Theorem 1 implies that $\mathcal{K}'_{\text{opt}}(\boldsymbol{0}) \neq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}(\boldsymbol{0})\big\}$. Since $\mathcal{K}_{\text{opt}}(\boldsymbol{0})$ is nonempty, $\boldsymbol{0} \in \{\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\}$. Thus, Problems ($P_2$) and ($P_4$) are not equivalent (equivalence fails for $\boldsymbol{z} = \boldsymbol{0}$). Conversely, suppose that $\exists\, \boldsymbol{z} \in \{\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y}) \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\}$ resulting in $\mathcal{K}'_{\text{opt}}(\boldsymbol{z}) \neq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}(\boldsymbol{z})\big\}$. Using last part of Theorem 1, we have $\{\boldsymbol{0}\} \subsetneq \big\{\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}} \;\big|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}_{\text{opt}}(\boldsymbol{z})\big\}$, which is possible only if $\boldsymbol{z} = \mathscr{S}(\boldsymbol{0}) = \boldsymbol{0}$. Now using (41) we get $\{\boldsymbol{0}\} \subsetneq \mathcal{K}' \bigcap \mathcal{N}(\mathscr{S}, 1)$.

Appendix C
Proof of Proposition 1

Problem (P_4) fails if and only if it admits more than one optimal solution.

Let $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$ and for the sake of contradiction suppose that $\boldsymbol{W}_1 \in \mathcal{K}'$ and $\boldsymbol{W}_2 \in \mathcal{K}'$ denote two solutions to Problem (P_4) for some observation $\boldsymbol{z}$, so that $(\boldsymbol{W}_1 - \boldsymbol{W}_2) \in \mathcal{M}$. Then, $\mathscr{S}(\boldsymbol{W}_1) = \mathscr{S}(\boldsymbol{W}_2)$ so that $(\boldsymbol{W}_1 - \boldsymbol{W}_2)$ is in the null

space of $\mathscr{S}$. But, $\text{rank}(\boldsymbol{W}_1 - \boldsymbol{W}_2) \leq \text{rank}(\boldsymbol{W}_1) + \text{rank}(\boldsymbol{W}_2) \leq 2$ so that we have $\boldsymbol{W}_1 - \boldsymbol{W}_2 = \boldsymbol{0}$ and Problem (P_4) has a unique solution.

Conversely, let Problem (P_4) have a unique solution for every observation $\boldsymbol{z} = \boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y})$. For the sake of contradiction, suppose that there is a matrix $\boldsymbol{Y}$ in $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$. Since $\boldsymbol{Y} \in \mathcal{M}$, $\exists \boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathcal{K}'$ such that $\boldsymbol{Y} = \boldsymbol{Y}_1 - \boldsymbol{Y}_2$. Further, $\boldsymbol{Y} \neq \boldsymbol{0}$ is in the null space of $\mathscr{S}$, so that $\boldsymbol{z} = \mathscr{S}(\boldsymbol{Y}_1) = \mathscr{S}(\boldsymbol{Y}_2)$ with $\boldsymbol{Y}_1 \neq \boldsymbol{Y}_2$ implying that $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are both valid solutions to Problem (P_4) for the observation $\boldsymbol{z}$. Since $\mathcal{K}' = \left\{\boldsymbol{x}\boldsymbol{y}^{\text{T}} \;\middle|\; (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{K}\right\}$, $\exists (\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in \mathcal{K}$ such that $\boldsymbol{S}(\boldsymbol{x}_1, \boldsymbol{y}_1) = \mathscr{S}(\boldsymbol{Y}_1)$ and $\boldsymbol{S}(\boldsymbol{x}_2, \boldsymbol{y}_2) = \mathscr{S}(\boldsymbol{Y}_2)$, so that $\boldsymbol{z} = \mathscr{S}(\boldsymbol{Y}_1) = \mathscr{S}(\boldsymbol{Y}_2)$ is a valid observation. This violates the unique solution assumption on Problem (P_4) for the valid observation $\boldsymbol{z}$. Hence $\mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$, completing the proof.

Appendix D
Proof of Theorem 2

Let $\boldsymbol{M}^* \in \mathcal{K}'$ be a solution to Problem (P_4) such that $\boldsymbol{M}^* \neq \boldsymbol{M}$. Since $\boldsymbol{M}$ is a valid solution to Problem (P_4), we have $\text{rank}(\boldsymbol{M}^*) = \text{rank}(\boldsymbol{M}) = 1$ and $\boldsymbol{X} = \boldsymbol{M} - \boldsymbol{M}^* \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$. If $\boldsymbol{M}^* = \sigma_* \boldsymbol{u}_* \boldsymbol{v}_*^{\text{T}}$, then $\mathcal{R}(\boldsymbol{X}) = \text{Span}(\boldsymbol{v}, \boldsymbol{v}_*)$ and $\mathcal{C}(\boldsymbol{X}) = \text{Span}(\boldsymbol{u}, \boldsymbol{u}_*)$. This contradicts the assumption that at least one of $\boldsymbol{u} \notin \mathcal{C}(\boldsymbol{X})$ or $\boldsymbol{v} \notin \mathcal{R}(\boldsymbol{X})$ is true and completes the proof.

Appendix E
Proof of Corollary 2

We start with the "if" part. For $\boldsymbol{M}$ to be identifiable, we need $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) > 1$ for every matrix $\boldsymbol{X} \neq \boldsymbol{M}$ in the null space of $\mathscr{S}(\cdot)$ that also satisfies $\boldsymbol{M} - \boldsymbol{X} \in \mathcal{K}'$. Since $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) \geq \text{rank}(\boldsymbol{X}) - \text{rank}(\boldsymbol{M}) = \text{rank}(\boldsymbol{X}) - 1$, it is sufficient to consider matrices $\boldsymbol{X}$ with $\text{rank}(\boldsymbol{X}) \leq 2$. Thus, for identifiability of $\boldsymbol{M}$, we need $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) > 1, \forall \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap (\boldsymbol{M} - \mathcal{K}') \setminus \{\boldsymbol{0}\}$. Using $\mathcal{N}(\mathscr{S}, 1) \bigcap \mathcal{M} = \{\boldsymbol{0}\}$ and $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap (\boldsymbol{M} - \mathcal{K}') \setminus \{\boldsymbol{0}\}$, we have $\boldsymbol{u} \in \mathcal{C}(\boldsymbol{X})$ and $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{X})$ and by assumption, we have $\sigma_1(\boldsymbol{X}) = \sigma_2(\boldsymbol{X})$. Let $\boldsymbol{X} = \sigma_* \boldsymbol{u}_1 \boldsymbol{v}_1^{\text{T}} + \sigma_* \boldsymbol{u}_2 \boldsymbol{v}_2^{\text{T}}$ and $\boldsymbol{u} = \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2$, $\boldsymbol{v} = \alpha_3 \boldsymbol{v}_1 + \alpha_4 \boldsymbol{v}_2$ for some $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}$ with $\alpha_1^2 + \alpha_2^2 = \alpha_3^2 + \alpha_4^2 = 1$. It is easy to check that $\boldsymbol{X}$ has the following equivalent singular value decompositions,

$$\boldsymbol{X} = \sigma_* \boldsymbol{u}_1 \boldsymbol{v}_1^{\text{T}} + \sigma_* \boldsymbol{u}_2 \boldsymbol{v}_2^{\text{T}} = \sigma_* (\alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2)(\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2)^{\text{T}} + \sigma_* (\alpha_2 \boldsymbol{u}_1 - \alpha_1 \boldsymbol{u}_2)(\alpha_2 \boldsymbol{v}_1 - \alpha_1 \boldsymbol{v}_2)^{\text{T}}. \tag{42}$$

Using the representations for $\boldsymbol{u}$ and $\boldsymbol{v}$, we have,

$$\boldsymbol{M} - \boldsymbol{X} = -\sigma_* (\alpha_2 \boldsymbol{u}_1 - \alpha_1 \boldsymbol{u}_2)(\alpha_2 \boldsymbol{v}_1 - \alpha_1 \boldsymbol{v}_2)^{\text{T}} + (\alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2)[(\sigma \alpha_3 - \sigma_* \alpha_1) \boldsymbol{v}_1 + (\sigma \alpha_4 - \sigma_* \alpha_2) \boldsymbol{v}_2]^{\text{T}}. \tag{43}$$

As the column vectors $\boldsymbol{u} = \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2$ and $\boldsymbol{u}' = \alpha_2 \boldsymbol{u}_1 - \alpha_1 \boldsymbol{u}_2$ on the right hand side of (43) are linearly independent, $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) = 1$ is possible if and only if every column of $\boldsymbol{M} - \boldsymbol{X}$ combines $\boldsymbol{u}$ and $\boldsymbol{u}'$ in the same ratio. This means that the row vectors on the r.h.s. of (43) are scalar multiples of each other. Thus, for $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) = 1$ it is necessary that

$$\frac{\sigma \alpha_3 - \sigma_* \alpha_1}{\alpha_2} = \frac{\sigma \alpha_4 - \sigma_* \alpha_2}{-\alpha_1} \tag{44a}$$

or equivalently,

$$\sigma \alpha_1 \alpha_3 + \sigma \alpha_2 \alpha_4 = \sigma_* \alpha_1^2 + \sigma_* \alpha_2^2 = \sigma_* \tag{44b}$$

which is not possible unless,

$$\alpha_1 \alpha_3 + \alpha_2 \alpha_4 = \frac{\sigma_*}{\sigma} > 0. \tag{44c}$$

So, $\alpha_1 \alpha_3 + \alpha_2 \alpha_4 \leq 0$ implies that $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) > 1$. As $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap (\boldsymbol{M} - \mathcal{K}') \setminus \{\boldsymbol{0}\}$ is arbitrary, $\boldsymbol{M}$ is identifiable by Problem (P_4).

Next we prove the "only if" part. Let $\boldsymbol{M}$ be identifiable and $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap (\boldsymbol{M} - \mathcal{K}') \setminus \{\boldsymbol{0}\}$ so that $\boldsymbol{M} - \boldsymbol{X}$ is feasible for Problem (P_4). As before, we have $\boldsymbol{u} \in \mathcal{C}(\boldsymbol{X})$, $\boldsymbol{v} \in \mathcal{R}(\boldsymbol{X})$ and $\sigma_1(\boldsymbol{X}) = \sigma_2(\boldsymbol{X})$. If $\boldsymbol{X} = \sigma_* \boldsymbol{u}_1 \boldsymbol{v}_1^{\text{T}} + \sigma_* \boldsymbol{u}_2 \boldsymbol{v}_2^{\text{T}}$, then $\boldsymbol{u} = \alpha_1 \boldsymbol{u}_1 + \alpha_2 \boldsymbol{u}_2$, $\boldsymbol{v} = \alpha_3 \boldsymbol{v}_1 + \alpha_4 \boldsymbol{v}_2$ for some $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}$ with $\alpha_1^2 + \alpha_2^2 = \alpha_3^2 + \alpha_4^2 = 1$. It is simple to check that (42) and (43) are valid. We shall now assume $\epsilon = \alpha_1 \alpha_3 + \alpha_2 \alpha_4 > 0$ and arrive at a contradiction. Since multiplying a matrix by a nonzero scalar does not change its row or column space and scales every nonzero singular value in the same ratio, we can take $\sigma_* = \sigma \epsilon$ without violating any assumptions on $\boldsymbol{X}$. Thus, $\alpha_1 \alpha_3 + \alpha_2 \alpha_4 = \sigma_* / \sigma$ and we have (44c) $\implies$ (44b) $\implies$ (44a) $\implies$ $\text{rank}(\boldsymbol{M} - \boldsymbol{X}) = 1$ (the last implication is due to (43)) thus contradicting the identifiability of $\boldsymbol{M}$.

Appendix F
Proof of Lemma 1

Using assumption (A1), we have

$$\text{E}\left[\|\boldsymbol{x}\|_2^2\right] = \text{E}\left[\boldsymbol{x}^{\text{T}} \boldsymbol{x}\right] = \text{E}\left[\text{Tr}\left(\boldsymbol{x}\boldsymbol{x}^{\text{T}}\right)\right] = \text{Tr}\left(\text{E}\left[\boldsymbol{x}\boldsymbol{x}^{\text{T}}\right]\right) = \text{Tr}(\mathbf{I}) = m. \tag{45}$$

Hence,

$$\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\Big] = \frac{1}{m}\,\mathrm{E}\Big[\|\boldsymbol{x}\|_2^2\Big]\,\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\Big] \tag{46a}$$

$$= \frac{1}{m}\,\mathrm{E}\Big[\|\boldsymbol{x}\|_2^2\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\Big] \tag{46b}$$

$$= \frac{1}{m}\,\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big\|_2^2\Big] \tag{46c}$$

$$= \frac{1}{m}\,\mathrm{E}\big[\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big] \tag{46d}$$

$$= \frac{1}{m}\,\mathrm{E}\big[\mathrm{Tr}\big(\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big)\big]$$

$$= \frac{1}{m}\,\mathrm{Tr}\big(\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\,\mathrm{E}\big[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]\big) \tag{46e}$$

$$= \frac{1}{m}\,\mathrm{Tr}\big(\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\mathbf{I}\big) \tag{46f}$$

$$\leq \frac{2}{m} \tag{46g}$$

where (46a) follows from (45), (46b) and (46c) are true because $\boldsymbol{u} = \boldsymbol{x}/\|\boldsymbol{x}\|_2$ and assumption (A3) implies independence of $\|\boldsymbol{x}\|_2$ and $\boldsymbol{u}$, (46d) is true since $\boldsymbol{P}^2 = \boldsymbol{P}$ for any projection matrix $\boldsymbol{P}$, (46e) is true since expectation operator commutes with trace and projection operators, (46f) follows from assumption (A1) and, (46g) is true since $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}$ is a matrix of rank at most two.

Finally, applying Markov inequality to the non-negative random variable $\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2$ and using the computed estimate of $\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\Big]$ from (46) gives

$$\Pr\Big(\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2 \geq 1-\delta\Big) \leq \frac{\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\Big]}{1-\delta} = \frac{2}{m(1-\delta)}. \tag{47}$$

We have thus established (14a). Using the exact same sequence of steps for the random vector $\boldsymbol{v}$ gives the bound in (14b).

Appendix G
Proof of Lemma 2

Notice that (46d)-(46g) in the proof of Lemma 1 in Appendix F does not use assumption (A3). Hence, reusing the same arguments we get

$$\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big\|_2^2\Big] = \mathrm{E}\big[\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big] \leq 2. \tag{48}$$

Thus, we have

$$\Pr\Big(\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2 \geq 1-\delta\Big) = \Pr\Big(\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big\|_2^2 \geq (1-\delta)\|\boldsymbol{x}\|_2^2\Big) \tag{49a}$$

$$\leq \Pr\Big(\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big\|_2^2 \geq (1-\delta)r_{\boldsymbol{x}}^2\Big) \tag{49b}$$

$$\leq \frac{\mathrm{E}\Big[\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big\|_2^2\Big]}{r_{\boldsymbol{x}}^2(1-\delta)} \tag{49c}$$

$$\leq \frac{2}{r_{\boldsymbol{x}}^2(1-\delta)} \tag{49d}$$

where (49a) is true since $\boldsymbol{u} = \boldsymbol{x}/\|\boldsymbol{x}\|_2$, (49b) holds because of assumption (A4), (49c) follows from applying Markov inequality to the non-negative random variable $\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\big\|_2^2$ and, (49d) follows from (48). Thus, the derivation (49) establishes (15a). Using the same sequence of steps for the random vector $\boldsymbol{v}$ gives the bound in (15b).

Appendix H
Proof of Theorem 3

For any constant $\delta \in (0,1)$, let $\mathcal{A}(\delta)$ denote the event that $\exists \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2)\bigcap\mathcal{M}\setminus\{\mathbf{0}\}$ satisfying both $\big\|\boldsymbol{u} - \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2 \leq \delta$ and $\big\|\boldsymbol{v} - \boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\big\|_2^2 \leq \delta$. We note that $\mathcal{A}(\delta)$ constitutes a non-decreasing sequence of sets as δ increases. Hence, using continuity of the probability measure from above we have,

$$\Pr(\mathcal{A}(0)) \leq \Pr(\mathcal{A}(\delta)) \tag{50a}$$

for any $\delta \in (0,1)$, and

$$\Pr(\mathcal{A}(0)) = \lim_{\delta\to 0}\Pr(\mathcal{A}(\delta)). \tag{50b}$$

Note that $\mathcal{A}(0)$ denotes the event that $\exists \boldsymbol{X} \in \mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\boldsymbol{0}\}$ satisfying both $\boldsymbol{u}\in\mathcal{C}(\boldsymbol{X})$ and $\boldsymbol{v}\in\mathcal{R}(\boldsymbol{X})$ which is a "hard" event. The event $\mathcal{A}(0)^{\mathrm{c}}$ corresponds precisely to the sufficient conditions of Theorem 2. Hence, it is sufficient to obtain an appropriate lower bound for $\Pr(\mathcal{A}(0)^{\mathrm{c}})$ to make our desired statement. Drawing inspiration from (50a) and (50b), we shall upper bound $\Pr(\mathcal{A}(0))$ by $\Pr(\mathcal{A}(\delta))$.

For any given $\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\boldsymbol{0}\}$ we have,

$$\begin{aligned}
&\Pr\Big(\big\|\boldsymbol{u}-\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\le\delta,\big\|\boldsymbol{v}-\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\big\|_2^2\le\delta\Big)\\
&=\Pr\Big(\|\boldsymbol{u}\|_2^2-\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\le\delta,\|\boldsymbol{v}\|_2^2-\big\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\big\|_2^2\le\delta\Big) &\text{(51a)}\\
&=\Pr\Big(\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\ge1-\delta,\big\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\big\|_2^2\ge1-\delta\Big) &\text{(51b)}\\
&=\Pr\Big(\big\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\ge1-\delta\Big)\Pr\Big(\big\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\big\|_2^2\ge1-\delta\Big) &\text{(51c)}\\
&\le\frac{4}{mn(1-\delta)^2} &\text{(51d)}
\end{aligned}$$

where (51a) is true because $\mathbf{I}-\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}$ (respectively $\mathbf{I}-\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}$) is the orthogonal projection matrix onto the orthogonal complement space of $\mathcal{C}(\boldsymbol{X})$ (respectively $\mathcal{R}(\boldsymbol{X})$), (51b) is true because we have $\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1$, (51c) is true by independence of $\boldsymbol{u}$ and $\boldsymbol{v}$, and (51d) comes from applying Lemma 1.

Next we employ union bounding over all $\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\boldsymbol{0}\}$ representing distinct pairs of column and row subspaces $(\mathcal{C}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{X}))$ to upper bound $\Pr(\mathcal{A}(\delta))$. We denote the number of these distinct pairs of $(\mathcal{C}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{X}))$ over $\boldsymbol{X}\in\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}\setminus\{\boldsymbol{0}\}$ by $f_{\mathscr{S},\mathcal{M}}(m,n)$.

Finally, using (51) we have

$$\begin{aligned}
\Pr(\mathcal{A}(\delta))&\le\sum_{(\mathcal{C}(\boldsymbol{X}),\mathcal{R}(\boldsymbol{X}))}\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{u}\right\|_2^2\le\delta,\left\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})^\perp}\boldsymbol{v}\right\|_2^2\le\delta\right) &\text{(52a)}\\
&=f_{\mathscr{S},\mathcal{M}}(m,n)\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{u}\right\|_2^2\le\delta,\left\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})^\perp}\boldsymbol{v}\right\|_2^2\le\delta\right)\\
&\le\frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{mn(1-\delta)^2} &\text{(52b)}
\end{aligned}$$

where (52a) is an union bounding step. Hence,

$$\begin{aligned}
\Pr(\mathcal{A}(0)^{\mathrm{c}})&=1-\Pr(\mathcal{A}(0))\\
&\ge1-\Pr(\mathcal{A}(\delta)) &\text{(53a)}\\
&\ge1-\frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{mn(1-\delta)^2} &\text{(53b)}\\
&\ge1-\frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{mn(1-\delta')} &\text{(53c)}
\end{aligned}$$

where (53a) is from (50a), (53b) is from (52) and $\delta'=1-(1-\delta)^2\in(0,1)$.

Appendix I
Proof of Corollary 3

The proof is essentially to that of Theorem 3 in Appendix H with one important difference: we use Lemma 2 instead of Lemma 1 when bounding the right hand side of (51c). This gives us the bound

$$\Pr\Big(\big\|\boldsymbol{u}-\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{u}\big\|_2^2\le\delta,\big\|\boldsymbol{v}-\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{v}\big\|_2^2\le\delta\Big)\le\frac{4}{r_{\boldsymbol{x}}^2(m)r_{\boldsymbol{y}}^2(n)(1-\delta)^2} \tag{54}$$

which leads to the bound

$$\Pr(\mathcal{A}(\delta))\le\frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{r_{\boldsymbol{x}}^2(m)r_{\boldsymbol{y}}^2(n)(1-\delta)^2} \tag{55}$$

in place of (52b). Finally,

$$\begin{aligned}
\Pr(\mathcal{A}(0)^{\mathrm{c}})&=1-\Pr(\mathcal{A}(0))\\
&\ge1-\Pr(\mathcal{A}(\delta)) &\text{(56a)}\\
&\ge1-\frac{4f_{\mathscr{S},\mathcal{M}}(m,n)}{r_{\boldsymbol{x}}^2(m)r_{\boldsymbol{y}}^2(n)(1-\delta)^2} &\text{(56b)}
\end{aligned}$$

$$\geq 1 - \frac{4 f_{\mathscr{S},\mathcal{M}}(m,n)}{r_{\boldsymbol{x}}^2(m) r_{\boldsymbol{y}}^2(n)(1-\delta')} \tag{56c}$$

where (56a) is from (50a), (56b) is from (55) and $\delta' = 1 - (1-\delta)^2 \in (0,1)$.

Appendix J
Proof of Lemma 3

This is a Chernoff-type bound. We set

$$Y = \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 - (1-\delta)\|\boldsymbol{x}\|_2^2 = \delta\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 - (1-\delta)\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{x}\right\|_2^2 \tag{57}$$

and compute the bound

$$\Pr(Y \geq 0) \leq \mathrm{E}[\exp(tY)] \tag{58}$$

that holds for all values of the parameter t for which the right hand side of (58) exists. Using properties of Gaussian random vectors under linear transforms, we have $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}$ and $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{x}$ as statistically independent Gaussian random vectors implying

$$\mathrm{E}\left[\exp\left(t\delta\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 - t(1-\delta)\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{x}\right\|_2^2\right)\right] = \mathrm{E}[\exp(t\delta Z_1)]\cdot\mathrm{E}[\exp(-t(1-\delta)Z_2)]. \tag{59}$$

with,

$$Z_1 = \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 \quad \text{and} \quad Z_2 = \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{x}\right\|_2^2. \tag{60}$$

Since $\mathcal{N}(\mathscr{S},1)\bigcap\mathcal{M} = \{\boldsymbol{0}\}$, both $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{R}(\boldsymbol{X})$ are two dimensional spaces. On rotating coordinates to the basis given by $\left\{\mathcal{C}(\boldsymbol{X}), \mathcal{C}(\boldsymbol{X})^\perp\right\}$, it can be seen that Z_1 is the sum of squares of two *i.i.d.* standard Gaussian random variables and hence has a χ^2 distribution with two DoF. By the same argument, Z_2 is a χ^2 distributed random variable with $(m-2)$ DoF. Recall that the moment generating function of a χ^2 distributed random variable Z with k DoF is given by

$$\mathrm{E}[\exp(tZ)] = (1-2t)^{-k/2}, \quad \forall t < 1/2. \tag{61}$$

Using (57), (58), (59) and (61) we have the bound

$$\begin{aligned}\Pr(Y \geq 0) &\leq (1-2t\delta)^{-1}(1-2t(1-\delta))^{-(m-2)/2} \\ &= \exp\left[-\left(\frac{m-2}{2}\right)\log(1-2t(1-\delta)) - \log(1-2t\delta)\right]\end{aligned} \tag{62}$$

which can be optimized over t. It can be verified by differentiation that the best bound is obtained for

$$t^* = \frac{m-2}{2\delta m} - \frac{1}{m(1-\delta)}. \tag{63}$$

Plugging this value of t into (62) and using (57) we get the desired result.

Appendix K
Proof of Lemma 4

This is also a Chernoff-type bound. Although the final results of Lemmas 3 and 4 look quite similar, we cannot reuse the manipulations in Appendix J for this proof and proceed by a slightly different route (also applicable to other subgaussian distributions) since the symmetric Bernoulli distribution does not share the rotational invariance property of the multivariate standard normal distribution. Let $\{\boldsymbol{c}_1, \boldsymbol{c}_2\}$ denote an orthonormal basis for $\mathcal{C}(\boldsymbol{X})$ and set

$$Y = \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 - (1-\delta)\|\boldsymbol{x}\|_2^2. \tag{64}$$

Notice that $\|\boldsymbol{x}\|_2 = \sqrt{m}$, so we have

$$\begin{aligned}\Pr(Y \geq 0) &= \Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 \geq m(1-\delta)\right) \\ &= \Pr\left(|\langle\boldsymbol{c}_1,\boldsymbol{x}\rangle|^2 + |\langle\boldsymbol{c}_2,\boldsymbol{x}\rangle|^2 \geq m(1-\delta)\right) \\ &\leq \Pr\left(\bigcup_{j=1,2}\left\{|\langle\boldsymbol{c}_j,\boldsymbol{x}\rangle|^2 \geq \frac{m}{2}(1-\delta)\right\}\right) \qquad (65a) \\ &\leq 2\Pr\left(|\langle\boldsymbol{c},\boldsymbol{x}\rangle|^2 \geq \frac{m}{2}(1-\delta)\right) \qquad (65b)\end{aligned}$$

$$= 2\Pr\left(|\langle \boldsymbol{c}, \boldsymbol{x}\rangle| \geq \sqrt{\frac{m}{2}(1-\delta)}\right)$$
$$= 4\Pr\left(\langle \boldsymbol{c}, \boldsymbol{x}\rangle \geq \sqrt{\frac{m}{2}(1-\delta)}\right) \tag{65c}$$
$$\leq 4\exp\left[\frac{t^2}{2} - t\sqrt{\frac{m}{2}(1-\delta)}\right] \tag{65d}$$

where (65a) and (65b) utilize elementary union bounds, $\boldsymbol{c}$ is a generic unit vector, (65c) uses the symmetry of the distribution of $\boldsymbol{x}$ about the origin, and (65d) is the Chernoff bounding step that utilizes the following computation:

$$\mathrm{E}[\exp(t\langle \boldsymbol{c}, \boldsymbol{x}\rangle)] = \mathrm{E}\left[\exp\left(\sum_{j=1}^{m} t x_j c_j\right)\right]$$
$$= \mathrm{E}\left[\prod_{j=1}^{m} \exp(t x_j c_j)\right]$$
$$= \prod_{j=1}^{m} \mathrm{E}[\exp(t x_j c_j)] \tag{66a}$$
$$= \prod_{j=1}^{m} \frac{e^{t c_j} + e^{-t c_j}}{2} \tag{66b}$$
$$= \prod_{j=1}^{m} \sum_{k=0}^{\infty} \frac{(t c_j)^{2k}}{(2k)!} \tag{66c}$$
$$< \prod_{j=1}^{m} \sum_{k=0}^{\infty} \frac{(t c_j)^{2k}}{2^k k!} \tag{66d}$$
$$= \prod_{j=1}^{m} \exp\left(t^2 c_j^2/2\right)$$
$$= \exp\left(\frac{t^2}{2}\sum_{j=1}^{m} c_j^2\right)$$
$$= \exp\left(\frac{t^2}{2}\right) \tag{66e}$$

where (66a) uses independence of elements of $\boldsymbol{x}$, (66b) is true because each element of $\boldsymbol{x}$ has a symmetric Bernoulli distribution, (66c) uses the series expansion of the exponential function, (66e) follows from $\|\boldsymbol{c}\|_2 = 1$ and (66d) is due to

$$(2k)! = 2^k \prod_{r=0}^{k-1}(2r+1) > 2^k \prod_{r=0}^{k-1}(r+1) = 2^k k!. \tag{67}$$

The bound in (65d) can be optimized over t with the optimum being achieved at

$$t^* = \sqrt{\frac{m}{2}(1-\delta)}. \tag{68}$$

Plugging this value of t into (65d) gives the desired result.

Appendix L
Proof of Lemma 5

Consider the norm $\|\cdot\|_{2,\infty}$ on $\mathbb{R}^{m\times 2}$ defined as

$$\|\boldsymbol{Y}\|_{2,\infty} = \max\{\|\boldsymbol{y}_1\|_2, \|\boldsymbol{y}_2\|_2\} \tag{69}$$

for all $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2] \in \mathbb{R}^{m\times 2}$. It is clear that $\mathcal{D}(m)$ is the unit ball $\left\{\boldsymbol{Y} \in \mathbb{R}^{m\times 2} \;\middle|\; \|\boldsymbol{Y}\|_{2,\infty} \leq 1\right\}$ of this norm, which is a convex body symmetric about the origin. Hence, using (21) we have the metric entropy of $\mathcal{D}(m)$ *w.r.t.* $\epsilon\mathcal{D}(m)$ as $2m\log\Theta(1/\epsilon)$. It is clear that $\mathcal{G}(m) = \left\{\boldsymbol{Y} \in \mathbb{R}^{m\times 2} \;\middle|\; \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{I}\right\} \subsetneq \mathcal{D}(m)$, implying that metric entropy of $\mathcal{G}(m)$ *w.r.t.* $\epsilon\mathcal{D}(m)$ is $\leq 2m\log\Theta(1/\epsilon)$.

Let $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2]$ and $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2]$ be two elements from $\mathcal{G}(m)$ such that $\boldsymbol{Y} - \boldsymbol{Z} \in \epsilon\mathcal{D}(m)$, and let $\boldsymbol{x} \in \mathbb{R}^m$ be arbitrary.

Then,

$$\begin{aligned}
\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Y})}\boldsymbol{x}\right\|_2 &= \sqrt{|\langle \boldsymbol{y}_1, \boldsymbol{x}\rangle|^2 + |\langle \boldsymbol{y}_2, \boldsymbol{x}\rangle|^2} \\
&= \sqrt{\sum_{j=1,2} |\langle \boldsymbol{z}_j, \boldsymbol{x}\rangle + \langle \boldsymbol{y}_j - \boldsymbol{z}_j, \boldsymbol{x}\rangle|^2} \\
&\leq \sqrt{\sum_{j=1,2} \left(|\langle \boldsymbol{z}_j, \boldsymbol{x}\rangle| + |\langle \boldsymbol{y}_j - \boldsymbol{z}_j, \boldsymbol{x}\rangle|\right)^2} && \text{(70a)} \\
&\leq \sqrt{\sum_{j=1,2} \left(|\langle \boldsymbol{z}_j, \boldsymbol{x}\rangle| + \epsilon\|\boldsymbol{x}\|_2\right)^2} && \text{(70b)} \\
&= \left\| \epsilon\|\boldsymbol{x}\|_2 \begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}|\langle \boldsymbol{z}_1, \boldsymbol{x}\rangle|\\|\langle \boldsymbol{z}_2, \boldsymbol{x}\rangle|\end{bmatrix} \right\|_2 \\
&\leq \sqrt{2}\epsilon\|\boldsymbol{x}\|_2 + \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z})}\boldsymbol{x}\right\|_2 && \text{(70c)}
\end{aligned}$$

where (70a) is due to $(x+y)^2 \leq (|x|+|y|)^2, \forall x, y \in \mathbb{R}$, (70b) is due to the Cauchy-Schwartz inequality and the bound $\|\boldsymbol{y}_j - \boldsymbol{z}_j\|_2 \leq \epsilon, j = 1, 2$ as $\boldsymbol{Y} - \boldsymbol{Z} \in \epsilon\mathcal{D}(m)$, and (70c) is due to the triangle inequality. Since $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are interchangeable in the derivation of (70c) and $\boldsymbol{x}$ is arbitrary, we immediately arrive at (22).

Appendix M
Proof of Theorem 4

We follow a proof strategy similar to that of Theorem 3. For any constant $\delta \in (0,1)$, let $\mathcal{B}_c(\delta)$ (respectively $\mathcal{B}_r(\delta)$) denote the event that $\exists \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ satisfying, $\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 \geq (1-\delta)\|\boldsymbol{x}\|_2^2$ (respectively $\left\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{y}\right\|_2^2 \geq (1-\delta)\|\boldsymbol{y}\|_2^2$), and let $\mathcal{A}(\delta)$ denote the event that $\exists \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ satisfying both $\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{x}\right\|_2^2 \geq (1-\delta)\|\boldsymbol{x}\|_2^2$ and $\left\|\boldsymbol{P}_{\mathcal{R}(\boldsymbol{X})}\boldsymbol{y}\right\|_2^2 \geq (1-\delta)\|\boldsymbol{y}\|_2^2$. We note that $\mathcal{A}(\delta)$ constitutes a non-decreasing sequence of sets as δ increases. Hence, using continuity of the probability measure from above we have,

$$\Pr(\mathcal{A}(0)) \leq \Pr(\mathcal{A}(\delta)) \tag{71a}$$

for any $\delta \in (0,1)$, and

$$\Pr(\mathcal{A}(0)) = \lim_{\delta \to 0} \Pr(\mathcal{A}(\delta)). \tag{71b}$$

Note that $\mathcal{A}(0)$ denotes the event that $\exists \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ satisfying both $\boldsymbol{x} \in \mathcal{C}(\boldsymbol{X})$ and $\boldsymbol{y} \in \mathcal{R}(\boldsymbol{X})$ which is a "hard" event. The event $\mathcal{A}(0)^{\mathrm{c}}$ corresponds precisely to the sufficient conditions of Theorem 2. Hence, it is sufficient to obtain an appropriate lower bound for $\Pr(\mathcal{A}(0)^{\mathrm{c}})$, or alternatively, upper bound $\Pr(\mathcal{A}(0))$ using (71a). It is straightforward to see that

$$\begin{aligned}
\Pr(\mathcal{A}(\delta)) &\leq \Pr\Big(\mathcal{B}_r(\delta) \bigcap \mathcal{B}_c(\delta)\Big) && \text{(72a)} \\
&= \Pr(\mathcal{B}_r(\delta)) \Pr(\mathcal{B}_c(\delta)), && \text{(72b)}
\end{aligned}$$

where (72a) is because $\mathcal{A}(\delta)$ happens only when $\mathcal{B}_r(\delta)$ and $\mathcal{B}_c(\delta)$ are caused by the same matrix $\boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$, and (72b) is due to mutual independence between $\boldsymbol{x}$ and $\boldsymbol{y}$.

We have $\boldsymbol{x}$ and $\boldsymbol{y}$ drawn component-wise *i.i.d.* from a symmetric Bernoulli distribution. For any given $\boldsymbol{Y} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}$ we have a bound on $\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Y})}\boldsymbol{x}\right\|_2 \geq \sqrt{1-\delta}\|\boldsymbol{x}\|_2\right)$ from Lemma 4. We focus on the union bounding step to compute $\Pr(\mathcal{B}_c(\delta))$. The proof of Lemma 5 assures us that as long as $\boldsymbol{Y}, \boldsymbol{Z} \in \mathcal{G}(m) \bigcap \{\mathcal{C}(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{N}(\mathscr{S}, 2) \bigcap \mathcal{M} \setminus \{\boldsymbol{0}\}\}$ are close enough, *i.e.* within the same $\epsilon\mathcal{D}(m)$ ball for some $1 > \epsilon \geq \epsilon_0 > 0$, we are guaranteed tight control over $\left|\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Y})}\boldsymbol{x}\right\|_2 - \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z})}\boldsymbol{x}\right\|_2\right|$ for any arbitrary $\boldsymbol{x}$. In fact, using (70) we have the bound

$$\Pr\Big(\exists \boldsymbol{Y} \in \boldsymbol{Z} + \epsilon\mathcal{D}(m), \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Y})}\boldsymbol{x}\right\|_2 \geq \sqrt{1-\delta}\|\boldsymbol{x}\|_2\Big) \leq \Pr\Big(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z})}\boldsymbol{x}\right\|_2 \geq \left(\sqrt{1-\delta} - \sqrt{2}\epsilon\right)\|\boldsymbol{x}\|_2\Big). \tag{73}$$

Letting $\boldsymbol{Z}_k \in \mathbb{R}^{m \times 2}$ denote the center of the k^{th} $\epsilon\mathcal{D}(m)$ ball we have k ranging from 1 to $\exp[p_c \log \Theta(1/\epsilon)]$. We thus have $\Pr(\mathcal{B}_c(\delta))$ upper bounded by

$$\begin{aligned}
&\sum_k \Pr\Big(\exists \boldsymbol{Y} \in \boldsymbol{Z}_k + \epsilon\mathcal{D}(m), \left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Y})}\boldsymbol{x}\right\|_2 \geq \sqrt{1-\delta}\|\boldsymbol{x}\|_2\Big) && \text{(74a)} \\
&\quad\leq \sum_k \Pr\Big(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z}_k)}\boldsymbol{x}\right\|_2 \geq \left(\sqrt{1-\delta} - \sqrt{2}\epsilon\right)\|\boldsymbol{x}\|_2\Big) && \text{(74b)} \\
&\quad\leq \exp\left[p_c \log \Theta\left(\frac{1}{\epsilon}\right)\right] \Pr\Big(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z})}\boldsymbol{x}\right\|_2 \geq \sqrt{1-\delta'}\|\boldsymbol{x}\|_2\Big) && \text{(74c)}
\end{aligned}$$

$$\leq 4\exp\left[p_c \log\Theta\left(\frac{1}{\epsilon}\right) - \frac{m(1-\delta')}{4}\right] \tag{74d}$$

where (74a) is from an elementary union bound, (74b) is from (73), (74c) uses

$$\delta' = 1 - \left(\sqrt{1-\delta} - \sqrt{2}\epsilon\right)^2, \tag{75}$$

with $\boldsymbol{Z}$ being generic, and (74d) is true due to Lemma 4.

Replicating a similar sequence of steps to bound $\Pr(\mathcal{B}_r(\delta))$, one readily obtains the bound

$$\Pr(\mathcal{B}_r(\delta)) \leq 4\exp\left[p_r \log\Theta\left(\frac{1}{\epsilon}\right) - \frac{n(1-\delta')}{4}\right] \tag{76}$$

with δ' given by (75). Hence, combining (71a), (72b), (74d) and (76) we get

$$\Pr(\mathcal{A}(0)) \leq \Pr(\mathcal{B}_r(\delta))\Pr(\mathcal{B}_c(\delta)) \leq 16\exp\left[(p_c+p_r)\log\Theta\left(\frac{1}{\epsilon}\right) - (m+n)\frac{1-\delta'}{4}\right] \tag{77}$$

which yields the desired bound for $\Pr(\mathcal{A}(0)^{\mathrm{c}})$ when $p = p_c + p_r$.

Appendix N
Proof of Theorem 5

We have $\boldsymbol{x}$ and $\boldsymbol{y}$ drawn component-wise *i.i.d.* from a standard Gaussian distribution. The proof is essentially similar to that of Theorem 4 with one important difference (beside replacing all occurrences of $\mathcal{N}(\mathscr{S},2)\bigcap\mathcal{M}$ by $\mathcal{N}(\mathscr{S},2)$ and ϵ assuming values in $(0,1)$): we use the bound given by Lemma 3 instead of Lemma 4 when evaluating $\Pr\left(\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{Z})}\boldsymbol{x}\right\|_2 \geq \sqrt{1-\delta'}\|\boldsymbol{x}\|_2\right)$ in (74c). This gives us the bounds

$$\Pr(\mathcal{B}_c(\delta')) \leq C'(m,\delta')\exp\left[p_c\log\Theta\left(\frac{1}{\epsilon}\right) - m\log\frac{1}{\sqrt{\delta'}}\right] \tag{78a}$$

and (analogously),

$$\Pr(\mathcal{B}_r(\delta')) \leq C'(n,\delta')\exp\left[p_r\log\Theta\left(\frac{1}{\epsilon}\right) - n\log\frac{1}{\sqrt{\delta'}}\right] \tag{78b}$$

where

$$\begin{aligned} C'(m,\delta') &= \exp\left[2\log m - \frac{2}{m} + 2 - \log\frac{2\delta'}{1-\delta'}\right] \\ &= 0.5\exp(2)\left(\frac{1-\delta'}{\delta'}\right)\exp\left[2\log m - \frac{2}{m}\right] = \left(\frac{1-\delta'}{\delta'}\right)\Theta(m^2), \end{aligned} \tag{79}$$

As in (77), we have

$$\begin{aligned} \Pr(\mathcal{A}(0)) &\leq \Pr(\mathcal{B}_r(\delta'))\Pr(\mathcal{B}_c(\delta')) \\ &\leq C'(m,\delta')C'(n,\delta')\exp\left[(p_r+p_c)\log\Theta\left(\frac{1}{\epsilon}\right)\right]\exp\left[-(m+n)\log\frac{1}{\sqrt{\delta'}}\right] \end{aligned} \tag{80}$$

which gives the desired bound, since $p = p_c + p_r$ and

$$C'(m,\delta')C'(n,\delta') = \left(\frac{1-\delta'}{\delta'}\right)^2\Theta(m^2)\Theta(n^2) = C(m,n,\delta'). \tag{81}$$

Appendix O
Proof of Proposition 2

Let $\boldsymbol{X}$ admit a factorization as in (34). Then,

$$\boldsymbol{X} = \underbrace{\begin{bmatrix}\boldsymbol{0} & \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}} \\ 0 & \boldsymbol{0}^{\mathrm{T}}\end{bmatrix}}_{\boldsymbol{X}_1} + \underbrace{\begin{bmatrix}\boldsymbol{0}^{\mathrm{T}} & 0 \\ -\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}} & \boldsymbol{0}\end{bmatrix}}_{\boldsymbol{X}_2} \tag{82}$$

and we see that the matrix $\boldsymbol{X}_2$ is obtained by shifting down the elements of the matrix $\boldsymbol{X}_1$ by one unit along the anti diagonals, and then flipping the sign of each element. Since the convolution operator $\mathscr{S}(\cdot)$ sums elements along the anti diagonals (see Fig. 1 for illustration), the representation of $\boldsymbol{X}$ as in (82) immediately implies that $\mathscr{S}(\boldsymbol{X}) = \boldsymbol{0}$. Since (34) implies that $\mathrm{rank}(\boldsymbol{X}) \leq 2$ so we have $\boldsymbol{X} \in \mathcal{N}(\mathscr{S},2)$.