# Recovery of Low-Rank Plus Compressed Sparse Matrices With Application to Unveiling Traffic Anomalies

Morteza Mardani, *Student Member, IEEE*, Gonzalo Mateos, *Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Given the noiseless superposition of a low-rank matrix plus the product of a known fat compression matrix times a sparse matrix, the goal of this paper is to establish deterministic conditions under which exact recovery of the low-rank and sparse components becomes possible. This fundamental identifiability issue arises with traffic anomaly detection in backbone networks, and subsumes compressed sensing as well as the timely low-rank plus sparse matrix recovery tasks encountered in matrix decomposition problems. Leveraging the ability of $\ell_1$ and nuclear norms to recover sparse and low-rank matrices, a convex program is formulated to estimate the unknowns. Analysis and simulations confirm that the said convex program can recover the unknowns for sufficiently low-rank and sparse enough components, along with a compression matrix possessing an isometry property when restricted to operate on sparse vectors. When the low-rank, sparse, and compression matrices are drawn from certain random ensembles, it is established that exact recovery is possible with high probability. First-order algorithms are developed to solve the nonsmooth convex optimization problem with provable iteration complexity guarantees. Insightful tests with synthetic and real network data corroborate the effectiveness of the novel approach in unveiling traffic anomalies across flows and time, and its ability to outperform existing alternatives.**

***Index Terms*—Convex optimization, identifiability, low rank, sparsity, traffic volume anomalies.**

## I. Introduction

Let $\mathbf{X}_0 \in \mathbb{R}^{L\times T}$ be a low-rank matrix [$r := \text{rank}(\mathbf{X}_0) \ll \min(L,T)$], and let $\mathbf{A}_0 \in \mathbb{R}^{F\times T}$ be sparse ($s := \|\mathbf{A}_0\|_0 \ll FT$, $\|\cdot\|_0$ counts the nonzero entries of its matrix argument). Given a compression matrix $\mathbf{R} \in \mathbb{R}^{L\times F}$ with $L \leq F$, and observations

$$\mathbf{Y} = \mathbf{X}_0 + \mathbf{R}\mathbf{A}_0 \tag{1}$$

this paper deals with the recovery of $\{\mathbf{X}_0, \mathbf{A}_0\}$. This task is of interest, e.g., to unveil anomalous flows in backbone networks [32], [36], [54], to reduce the data acquisition time in cardiac magnetic resonance imaging (MRI) [25], [26], or, to separate singing voice from its music accompaniment [29], [46]; see also Section II on motivating applications. In addition, this fundamental problem is met at the crossroads of compressive sampling (CS), and the timely low-rank-plus-sparse matrix decompositions.

In the absence of the low-rank component ($\mathbf{X}_0 = \mathbf{0}_{L\times T}$), one is left with an under-determined sparse signal recovery problem; see, e.g., [15], [43] and the tutorial account [16]. When $\mathbf{Y} = \mathbf{X}_0 + \mathbf{A}_0$, the formulation boils down to principal component pursuit (PCP), also referred to as robust principal component analysis (PCA) [11], [17], [18], [22]. For this idealized noise-free setting, sufficient conditions for exact recovery are available for both of the aforementioned special cases; see also [18] for state-of-the-art PCP recovery guarantees, even valid when only a subset of $\mathbf{Y}$'s entries are observed. However, the superposition of a low-rank and a *compressed* sparse matrix in (1) further challenges identifiability of $\{\mathbf{X}_0, \mathbf{A}_0\}$. Along these lines, the *compressive* PCP formulation in [51] aims at recovering a target matrix that is a superposition of low-rank and sparse components, from a (small) set of linear measurements; see also [2] for a related approach. In the presence of "dense" noise, stable reconstruction of the low-rank and sparse matrix components is possible via PCP [53], [55]. Earlier efforts dealing with the recovery of sparse vectors in noise led to similar performance guarantees; see, e.g., [7] and references therein. Even when $\mathbf{X}_0$ is nonzero, one could envision a CS variant where the measurements are corrupted with correlated (low-rank) noise [19]. Last but not least, when $\mathbf{A}_0 = \mathbf{0}_{F\times T}$ and $\mathbf{Y}$ is noisy, the recovery of $\mathbf{X}_0$ subject to a rank constraint is nothing else than PCA—arguably, the workhorse of high-dimensional data analysis [31].

The main contribution of this paper is to establish that given $\mathbf{Y}$ and $\mathbf{R}$ in (1), for small enough $r$ and $s$, one can *exactly* recover $\{\mathbf{X}_0, \mathbf{A}_0\}$ by solving the nonsmooth *convex* optimization problem

$$\text{(P1)} \quad \min_{\{\mathbf{X},\mathbf{A}\}} \ \|\mathbf{X}\|_* + \lambda\|\mathbf{A}\|_1, \quad \text{s. t.} \quad \mathbf{Y} = \mathbf{X} + \mathbf{R}\mathbf{A}$$

where $\lambda \geq 0$ is a tuning parameter; $\|\mathbf{X}\|_* := \sum_i \sigma_i(\mathbf{X})$ is the nuclear norm of $\mathbf{X}$ ($\sigma_i$ stands for the $i$th singular value); and, $\|\mathbf{X}\|_1 := \sum_{i,j} |x_{ij}|$ denotes the $\ell_1$-norm. The aforementioned norms are convex surrogates to the rank and $\ell_0$-norm, respectively, which albeit natural as criteria they are NP-hard to optimize [20], [40]. Recently, a greedy algorithm for recovering low-rank and sparse matrices from compressive measurements

Manuscript received April 29, 2012; revised December 07, 2012; accepted April 04, 2013. Date of publication April 12, 2013; date of current version July 10, 2013. This work was supported in part by the NSF ECCS Grant No. 1202135 and in part by the NSF AST Grant No. 1247885. This paper was presented in part at the 2012 IEEE Statistical Signal Processing Workshop.

The authors are with the Department of Electrical and Computer Engineering and the Digital Technology Center, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: morteza@umn.edu; mate0058@umn.edu; georgios@umn.edu).

Communicated by M. Elad, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2257913

was put forth in [50]. However, convergence of the algorithm and its error performance are only assessed via numerical simulations. A recursive online algorithm can be found in [19], which attains good performance in practice but does not offer theoretical guarantees; see also [38].

A *deterministic* approach along the lines of [17] is adopted first to derive conditions under which (1) is locally identifiable (see Section III). Introducing a notion of incoherence between the additive components $\mathbf{X}_0$ and $\mathbf{RA}_0$, and resorting to the restricted isometry constants (RICs) of $\mathbf{R}$ [15], sufficient conditions are obtained to ensure that (P1) succeeds in exactly recovering the unknowns (see Section IV-A). Intuitively, the results here assert that if $r$ and $s$ are sufficiently small, the nonzero entries of $\mathbf{A}_0$ are sufficiently spread out, and subsets of columns of $\mathbf{R}$ behave as isometries, then (P1) exactly recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$. As a byproduct, recovery results for PCP and CS are also obtained by specializing the aforesaid conditions accordingly (see Section IV-B). However, these induced recovery guarantees are weaker than those recently obtained for PCP and CS by relying on state-of-the-art analysis techniques tailored to these specific problems; see, e.g., [18], [43], and references therein. The proof of the main result builds on Lagrangian duality theory [5], [10], to first derive conditions under which $\{\mathbf{X}_0, \mathbf{A}_0\}$ is the *unique* optimal solution of (P1) (see Section V-A). In a nutshell, satisfaction of the optimality conditions is tantamount to the existence of a valid dual certificate. Stemming from the unique challenges introduced by $\mathbf{R}$, the dual certificate construction procedure of Section V-B is markedly distinct from the direct sum approach in [17], and the (random) golfing scheme of [11]. Section VI shows that low-rank, sparse, and compression matrices drawn from certain random ensembles satisfy the sufficient conditions for exact recovery with high probability.

Two batch iterative algorithms for solving (P1) are developed in Section VII, based on the accelerated proximal gradient (APG) method [4], [34], [41], [42], and the alternating-direction method of multipliers (AD-MoM) [6], [10]. Decentralized and online algorithms were put forth in the companion papers [37] and [38]. These are useful when rows of $\mathbf{Y}$ are distributed over a network, and for real-time processing of streaming data (columns of $\mathbf{Y}$), respectively. Numerical tests corroborate the exact recovery claims, and the effectiveness of (P1) in unveiling traffic volume anomalies from real network data (see Section VIII). While the obtained sufficient conditions for exact recovery may be violated in the anomaly detection context of Section II-A, the encouraging results obtained in Section VIII-B suggest that there is room for improving these conditions. Section IX concludes this paper with a summary and a discussion of limitations, possible extensions, and interesting future directions. Technical details are deferred to the Appendix.

### A. Notational Conventions

Bold uppercase (lowercase) letters will denote matrices (column vectors), and calligraphic letters will denote sets. Operators $(\cdot)'$, $(\cdot)^{\dagger}$, $\mathrm{tr}(\cdot)$, $\mathrm{vec}(\cdot)$, $\mathrm{diag}(\cdot)$, $\lambda_{\max}(\cdot)$, $\sigma_{\min}(\cdot)$, and $\otimes$ will denote transposition, matrix pseudoinverse, matrix trace, matrix vectorization, diagonal matrix, spectral radius, minimum singular value, and Kronecker product, respectively; $|\cdot|$ will be used for the cardinality of a set and the magnitude of a scalar. The $n \times n$ identity matrix will be represented by $\mathbf{I}_n$ and its $i$th column by $\mathbf{e}_i$, while $\mathbf{0}_n$ denotes the $n \times 1$ vector of all zeros, and $\mathbf{0}_{n\times p} := \mathbf{0}_n\mathbf{0}'_p$. The $\ell_q$-norm of vector $\mathbf{x} \in \mathbb{R}^p$ is $\|\mathbf{x}\|_q := \left(\sum_{i=1}^{p} |x_i|^q\right)^{1/q}$ for $q \geq 1$. For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, define the trace inner product $\langle \mathbf{A}, \mathbf{B}\rangle := \mathrm{tr}(\mathbf{A}'\mathbf{B})$. Also, recall that $\|\mathbf{A}\|_F := \sqrt{\mathrm{tr}\,(\mathbf{AA}')}$ is the Frobenious norm, $\|\mathbf{A}\|_1 := \sum_{i,j} |a_{ij}|$ is the $\ell_1$-norm, $\|\mathbf{A}\|_\infty := \max_{i,j} |a_{ij}|$ is the $\ell_\infty$-norm, and $\|\mathbf{A}\|_* := \sum_i \sigma_i(\mathbf{A})$ is the nuclear norm. In addition, $\|\mathbf{A}\|_{1,1} := \max_{\|\mathbf{x}\|_1=1} \|\mathbf{Ax}\|_1 = \max_i \|\mathbf{e}'_i\mathbf{A}\|_1$ denotes the induced $\ell_1$-norm, and likewise for the induced $\ell_\infty$-norm, $\|\mathbf{A}\|_{\infty,\infty} := \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{Ax}\|_\infty = \max_i \|\mathbf{Ae}_i\|_1$. For the linear operator $\mathcal{A}$, define the operator norm $\|\mathcal{A}\| := \max_{\|\mathbf{X}\|_F=1} \|\mathcal{A}(\mathbf{X})\|_F$, which subsumes the spectral norm $\|\mathbf{A}\| := \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|$. Define also the support set $\mathrm{supp}(\mathbf{A}) := \{(i,j) : a_{ij} \neq 0\}$. The indicator function $\mathbb{1}_{\{a=b\}}$ equals one when $a = b$, and zero otherwise.

## II. Applications

This section outlines several application domains that involve decomposing a data matrix as in (1).

### A. Unveiling Network Anomalies Via Sparsity and Low Rank

In the backbone of large-scale networks, origin-to-destination (OD) traffic flows experience abrupt changes which can result in congestion and limit the quality of service provisioning of the end users. These so-termed traffic volume anomalies can be due to external sources such as network failures, denial of service attacks, or, intruders hijacking the network services [32], [48], [54]. Unveiling such anomalies is a crucial task toward engineering network traffic. This is a challenging task, however, since the available data are usually high-dimensional noisy link-load measurements, which comprise the superposition of *unobservable* OD flows as explained next.

Consider a backbone network with topology represented by the directed graph $G(\mathcal{N}, \mathcal{L})$, where $\mathcal{L}$ and $\mathcal{N}$ denote the set of links and nodes (routers) of cardinality $|\mathcal{L}| = L$ and $|\mathcal{N}| = N$, respectively. The network transports $F$ end-to-end flows associated with specific OD pairs. For backbone networks, the number of network layer flows is typically much larger than the number of physical links ($F \gg L$). Single-path routing is considered here to send the traffic flow from a source to its intended destination. Accordingly, for a particular flow, multiple links connecting the corresponding OD pair are chosen to carry the traffic. Sparing details that can be found in [36], the traffic $\mathbf{Y} := [y_{l,t}] \in \mathbb{R}^{L\times T}$ carried over links $l \in \mathcal{L}$ and measured at time instants $t \in [1, T]$ can be compactly expressed as

$$\mathbf{Y} = \mathbf{R}\,(\mathbf{Z} + \mathbf{A}) + \mathbf{E} \tag{2}$$

where the fat routing matrix $\mathbf{R} := [r_{\ell,f}] \in \{0,1\}^{L\times F}$ is fixed and given, $\mathbf{Z} := [z_{f,t}]$ denotes the unknown “clean” traffic flows over the time horizon of interest, $\mathbf{A} := [a_{f,t}]$ collects the traffic volume anomalies across flows and time, and $\mathbf{E} := [e_{l,t}]$ captures measurement errors.

Common temporal patterns among the traffic flows in addition to their periodic behavior, render most rows (respectively columns) of $\mathbf{Z}$ linearly dependent, and thus $\mathbf{Z}$ typically has low rank [32], [44]. Anomalies are expected to occur sporadically over time, and only last for short periods relative to the (possibly long) measurement interval $[1, T]$. In addition, only a small fraction of the flows are supposed to be anomalous at any given time instant. This renders the anomaly matrix $\mathbf{A}$ sparse across rows and columns. Given link measurements $\mathbf{Y}$ and the routing matrix $\mathbf{R}$, the goal is to estimate $\mathbf{A}$ by capitalizing on the sparsity of $\mathbf{A}$ and the low-rank property of $\mathbf{Z}$. Since the primary goal is to recover $\mathbf{A}$, define $\mathbf{X} := \mathbf{R}\mathbf{Z}$ which inherits the low-rank property from $\mathbf{Z}$, and consider

$$\mathbf{Y} = \mathbf{X} + \mathbf{R}\mathbf{A} + \mathbf{E} \tag{3}$$

which is identical to (1) modulo small measurement errors in $\mathbf{E} \in \mathbb{R}^{L \times T}$. If $\mathbf{E} = \mathbf{0}_{L \times T}$, then (P1) can be used to unveil network anomalies, whereas the algorithm outlined in Section VII-A is more suitable for the noisy setting.

By adopting the model (3), one is neglecting the structure $\mathbf{X} := \mathbf{R}\mathbf{Z}$. However, it is otherwise not clear how could one efficiently estimate $\mathbf{Z}$ and $\mathbf{A}$ from measurements as in (2), which is a more difficult problem. The compressive PCP approach [51] deals with the recovery of $\{\mathbf{Z}, \mathbf{A}\}$ from measurements $\tilde{\mathbf{Y}} = \mathcal{P}_Q(\mathbf{Z} + \mathbf{A}) \in \mathbb{R}^{F \times T}$, where $\mathcal{P}_Q(\cdot)$ denotes orthogonal projection onto a linear subspace $Q \subseteq \mathbb{R}^{F \times T}$. Note that compressive PCP cannot be adopted here since it requires $\{\mathbf{Z}, \mathbf{A}\}$ to be sufficiently incoherent with the orthogonal subspace $Q^{\perp}$, a condition which is violated in (2) since $Q^{\perp}$ is the nullspace of the fat compression matrix $\mathbf{R}$.

### B. Dynamic Magnetic Resonance Imaging

As a result of the existing limitations in MRI data-acquisition time, respiratory motions can severely degrade the quality of MRI. Consequently, this can result in, e.g., dose-delivery errors for patients subjected to radiation therapy [52]. *Dynamic* MRI aims at resolving the variations of the imaged object by reconstructing a temporal series of "ground truth" images [35]. As an illustrative example, consider cardiac MRI which nowadays serves as a major imaging modality for noninvasive diagnosis of heart diseases in clinic practice [24]. A critical specification of cardiac MRI is the simultaneous realization of higher spatial and temporal resolution. This, in turn, necessitates longer data-acquisition periods, which are however limited by the patient's breath-holding time. Inspired by the low intrinsic-dimensionality of (cardiac) MRI images [25], devising efficient techniques to reduce the acquisition time for a prescribed image quality becomes an important issue.

Consider each "ground truth" cardiac snapshot as a piecewise-constantly discretized image of $P$ pixels. Each image can be modeled as a superposition of a *background* component and a *motion* component [25], [26]. The background component refers to the temporally stationary or slowly varying part of the acquired images. Moreover, the motion component captures the rapidly changing pixels due to heart beating. The spatial structure of the heart has motivated the adoption of models involving a (possibly learnt and overcomplete) dictionary, under which the motion component admits a sparse representation based on few atoms (columns) of this dictionary [25], [26]. Let $\mathbf{x}_t \in \mathbb{R}^P$ denote the background component of the dynamic MRI frame acquired at time $t$, and let $\mathbf{D}\mathbf{a}_t \in \mathbb{R}^P$ denote the motion component, where $\mathbf{D} \in \mathbb{R}^{P \times F}$ is a given overcomplete dictionary, and $\mathbf{a}_t$ a sparse vector of coefficients. The MRI acquisition procedure entails measuring Fourier coefficients of the image, and only a subset of size $L \leq P$ of Fourier coefficients is sampled to reduce the data acquisition time. Accordingly, the partial FFT matrix $\mathbf{\Psi} \in \mathbb{R}^{L \times P}$ containing a row-subset of cardinality $L$ of the full FFT $P \times P$ matrix maps the image to a subset of its Fourier coefficients. The scanned temporal sequence of images in the frequency domain can thus be modeled as

$$\mathbf{y}_t = \mathbf{\Psi}(\mathbf{z}_t + \mathbf{D}\mathbf{a}_t) + \mathbf{v}_t, \quad t = 1, \ldots, T \tag{4}$$

where $\mathbf{v}_t$ accounts for modeling and measurement errors. Collect the components $\{\mathbf{x}_t := \mathbf{\Psi}\mathbf{z}_t\}_{t=1}^{T}$ and $\{\mathbf{a}_t\}_{t=1}^{T}$ as columns of the matrices $\mathbf{X}$ and $\mathbf{A}$, respectively, and recognize that (4) boils down to (1) upon defining $\mathbf{R} := \mathbf{\Psi}\mathbf{D}$. Notice that it suffices to estimate $\mathbf{X}$ (rather that $\mathbf{Z}$), since in cardiac MRI the main objective is to reconstruct the motion component $\mathbf{DA}$, which offers valuable information to physicians about possible heart diseases. By the very definition of background component, the sought matrix $\mathbf{X}$ is low rank. Also, $\mathbf{A}$ is sparse by construction of the dictionary $\mathbf{D}$. All in all, adopting (P1) to recover $\mathbf{A}$ and subsequently the motion component $\mathbf{DA}$ is well motivated.

### C. Face Recognition

Accurately estimating the low-dimensional subspace of a human's facial images is an important task in computer vision, with application to face recognition [3]. In this context, a robust approach is needed since facial images in the training set tend to be exposed to different illuminations, and typically suffer from specularities as well as self-shadowing (e.g., around the nose and eyes' areas). Similar to the dynamic MRI setup, a reasonable model represents each image as the superposition of a background (shadow-free face) component $\mathbf{X}$ which has low rank, and the error (shadow) component which is highly structured and localized. Model (1) is naturally aligned with this decomposition, upon learning a (possibly overcomplete) dictionary $\mathbf{D} := \mathbf{R}$ under which the error component $\mathbf{A}$ is sparsely represented. While PCP has been adopted in [11] to remove shadows and specularities from face images, (1) offers a more general alternative. This is because PCP presumes the sparse errors are independently scattered across the face image. However, this assumption neglects the fact that shadows and specularities usually contain certain spatial structure, which can be better modeled via a suitably learned dictionary of atoms.

### D. Separation of Singing Voice From Its Music Accompaniment

Separation of singing voice from its music accompaniment has wide applicability in areas such as automatic lyrics recognition and alignment, singer identification, and music information retrieval [33]. Even though this is an effortless task for the human auditory system, it is difficult for machines [29]. Let $\mathbf{Y}$ denote the spectrogram of a given song, which can be naturally modeled as the superposition of music plus singing-voice components. Due to the repetitious nature of music accompaniment, the music component $\mathbf{X}$ has low rank [29], [46]. In contrast, the singing voice exhibits higher variability, but as it is customary for speech signals [30], it can be reasonably assumed sparsely expressible over a proper dictionary $\mathbf{D} := \mathbf{R}$ of sounds. In a nutshell, (P1) can be adopted to carry out this decomposition task, while incorporating nonnegativity constraints on the matrix components is a natural extension since the spectrogram is inherently nonnegative [46].

## III. Local Identifiability

The first issue to address is model identifiability, meaning that there are *unique* low-rank and sparse matrices satisfying (1). If there exist multiple decompositions of $\mathbf{Y}$ into $\mathbf{X}+\mathbf{RA}$ with low-rank $\mathbf{X}$ and sparse $\mathbf{A}$, there is no hope of recovering $\{\mathbf{X}_0, \mathbf{A}_0\}$ from the data. For instance, if the null space of the fat matrix $\mathbf{R}$ contains sparse matrices, there may exist a sparse perturbation $\mathbf{H}$ such that $\mathbf{A}_0+\mathbf{H}$ is still sparse and $\{\mathbf{X}_0, \mathbf{A}_0+\mathbf{H}\}$ is a legitimate solution. Another problematic case arises when there is a sparse perturbation $\mathbf{H}$ such that $\mathbf{RH}$ is spanned by the row or column spaces of $\mathbf{X}_0$. Then, $\mathbf{X}_0+\mathbf{RH}$ has the same rank as $\mathbf{X}_0$ and $\mathbf{A}_0-\mathbf{H}$ may still be sparse. As a result, one may pick $\{\mathbf{X}_0+\mathbf{RH}, \mathbf{A}_0-\mathbf{H}\}$ as another valid solution. Dealing with such identifiability issues is the subject of this section.

Let $\mathbf{U\Sigma V}'$ denote the singular value decomposition (SVD) of $\mathbf{X}_0$, and consider the subspaces: s1) $\Phi(\mathbf{X}_0) := \{\mathbf{Z} \in \mathbb{R}^{L\times T} : \mathbf{Z} = \mathbf{UW}_1' + \mathbf{W}_2\mathbf{V}',\ \mathbf{W}_1 \in \mathbb{R}^{T\times r},\ \mathbf{W}_2 \in \mathbb{R}^{L\times r}\}$ of the span of all matrices with either the same column space or row space as $\mathbf{X}_0$; s2) $\Omega(\mathbf{A}_0) := \{\mathbf{H} \in \mathbb{R}^{F\times T} : \text{supp}(\mathbf{H}) \subseteq \text{supp}(\mathbf{A}_0)\}$ of matrices in $\mathbb{R}^{F\times T}$ with support contained in the support of $\mathbf{A}_0$; and s3) $\Omega_R(\mathbf{A}_0) := \{\mathbf{Z} \in \mathbb{R}^{L\times T} : \mathbf{Z} = \mathbf{RH},\ \mathbf{H} \in \Omega(\mathbf{A}_0)\}$. For notational brevity, s1)–s3) will be henceforth denoted as $\{\Phi, \Omega, \Omega_R\}$. Noteworthy properties of these subspaces are: i) both $\Phi$ and $\Omega_R \subset \mathbb{R}^{L\times T}$, hence it is possible to directly compare elements from them; ii) $\mathbf{X}_0 \in \Phi$ and $\mathbf{RA}_0 \in \Omega_R$; and iii) if $\mathbf{Z} \in \Phi^\perp$ is added to $\mathbf{X}_0$, then $\text{rank}(\mathbf{Z}+\mathbf{X}_0) > r$.

For now, assume that the subspaces $\Omega_R$ and $\Phi$ are also known. This extra information helps identifiability of (1), because potentially troublesome solutions $\{\mathbf{X}_0+\mathbf{RH}, \mathbf{A}_0-\mathbf{H}\}$ are limited to a restricted class. If $\mathbf{X}_0+\mathbf{RH} \notin \Phi$ or $\mathbf{A}_0-\mathbf{H} \notin \Omega$, that candidate solution is not admissible since it is known a priori that $\mathbf{A}_0 \in \Omega$ and $\mathbf{X}_0 \in \Phi$. Under these assumptions, the following lemma puts forth the necessary and sufficient conditions guaranteeing the existence of a *unique* pair of matrices $\{\mathbf{X}_0 \in \Phi, \mathbf{A}_0 \in \Omega_R\}$, such that $\mathbf{Y}$ can be decomposed according to (1) – a notion known as *local identifiability* [11], [17].

*Lemma 1:* Given subspaces $\{\Phi, \Omega, \Omega_R\}$ and matrices $\{\mathbf{Y}, \mathbf{R}\}$, there is a unique pair $\{\mathbf{X}_0 \in \Phi, \mathbf{A}_0 \in \Omega_R\}$ such that $\mathbf{Y} = \mathbf{X}_0 + \mathbf{RA}_0$ if and only if $\Phi \cap \Omega_R = \{\mathbf{0}_{L\times T}\}$, and $\mathbf{RH} \neq \mathbf{0}_{L\times T}, \forall \mathbf{H} \in \Omega\backslash\{\mathbf{0}_{F\times T}\}$.

*Proof:* Since by definition $\mathbf{X}_0 \in \Phi$ and $\mathbf{A}_0 \in \Omega$, one can represent every element in the *subspaces* $\Phi$ and $\Omega_R$ as $\mathbf{X}_0+\mathbf{Z}_1$ and $\mathbf{RA}_0+\mathbf{Z}_2$, respectively, where $\mathbf{Z}_1 \in \Phi$ and $\mathbf{Z}_2 \in \Omega_R$. Assume that $\Phi \cap \Omega_R = \{\mathbf{0}_{L\times T}\}$, and suppose by contradiction that there exist *nonzero* perturbations $\{\mathbf{Z}_1, \mathbf{Z}_2\}$ such that $\mathbf{Y} = \mathbf{X}_0+\mathbf{Z}_1+\mathbf{RA}_0+\mathbf{Z}_2$. Then, $\mathbf{Z}_1+\mathbf{Z}_2 = \mathbf{0}_{L\times T}$, meaning that $\mathbf{Z}_1$ and $\mathbf{Z}_2$ belong to the same subspace, which contradicts the assumption. Conversely, suppose there exists a nonzero $\mathbf{Z} \in \Omega_R \cap \Phi$. Clearly, $\{\mathbf{X}_0+\mathbf{Z}, \mathbf{RA}_0-\mathbf{Z}\}$ is a feasible solution where $\mathbf{X}_0+\mathbf{Z} \in \Phi$ and $\mathbf{RA}_0-\mathbf{Z} \in \Omega_R$. This contradicts the uniqueness assumption. In addition, the condition $\mathbf{RH} \neq \mathbf{0}, \mathbf{H} \in \Omega\backslash\{\mathbf{0}_{L\times T}\}$ ensures that $\mathbf{Z} = \mathbf{0}_{L\times T} \in \Phi\cap\Omega_R$ only when $\mathbf{Z} = \mathbf{RH} = \mathbf{0}_{L\times T}$ for $\mathbf{H} = \mathbf{0}_{F\times T}$. ■

In words, (1) is locally identifiable if and only if the subspaces $\Phi$ and $\Omega_R$ intersect transversally, and the sparse matrices in $\Omega$ are not annihilated by $\mathbf{R}$. This last condition is unique to the setting here and is not present in [11] or [17].

*Remark 1 (Orthogonal Projection Operators):* Operator $\mathcal{P}_\Omega(\mathbf{X})$ ($\mathcal{P}_{\Omega^\perp}(\mathbf{X})$) denotes the orthogonal projection of $\mathbf{X}$ onto the subspace $\Omega$ (orthogonal complement $\Omega^\perp$). It simply sets those elements of $\mathbf{X}$ not in $\text{supp}(\mathbf{A}_0)$ to zero. Likewise, $\mathcal{P}_\Phi(\mathbf{X})$ ($\mathcal{P}_{\Phi^\perp}(\mathbf{X})$) denotes the orthogonal projection of $\mathbf{X}$ onto the subspace $\Phi$ (orthogonal complement $\Phi^\perp$). Let $\mathbf{P}_U := \mathbf{UU}'$ and $\mathbf{P}_V := \mathbf{VV}'$ denote, respectively, projection onto the column and row spaces of $\mathbf{X}_0$. It can be shown that $\mathcal{P}_\Phi(\mathbf{X}) = \mathbf{P}_U\mathbf{X}+\mathbf{XP}_V-\mathbf{P}_U\mathbf{XP}_V$, while the projection onto the complement subspace is $\mathcal{P}_{\Phi^\perp}(\mathbf{X}) = (\mathbf{I}-\mathbf{P}_U)\mathbf{X}(\mathbf{I}-\mathbf{P}_V)$. In addition, the following identities

$$\langle \mathcal{P}_\Phi(\mathbf{X}), \mathcal{P}_\Phi(\mathbf{Y})\rangle = \langle \mathcal{P}_\Phi(\mathbf{X}), \mathbf{Y}\rangle = \langle \mathbf{X}, \mathcal{P}_\Phi(\mathbf{Y})\rangle \tag{5}$$

of orthogonal projection operators, such as $\mathcal{P}_\Phi(\cdot)$, will be invoked throughout this paper.

### A. Incoherence Measures

Building on Lemma 1, alternative sufficient conditions are derived here to ensure local identifiability. To quantify the overlap between $\Phi$ and $\Omega_R$, consider the *incoherence* parameter

$$\mu(\Omega_R, \Phi) = \max_{\mathbf{Z}\in\Omega_R\backslash\{\mathbf{0}\}} \frac{\|\mathcal{P}_\Phi(\mathbf{Z})\|_F}{\|\mathbf{Z}\|_F} \tag{6}$$

for which it holds that $\mu(\Omega_R, \Phi) \in [0,1]$. The lower bound is achieved when $\Phi$ and $\Omega_R$ are orthogonal, while the upper bound is attained when $\Phi \cap \Omega_R$ contains a nonzero element. Assuming $\Phi \cap \Omega_R = \{\mathbf{0}_{L\times T}\}$, then $\mu(\Omega_R, \Phi) < 1$ represents the cosine of the angle between $\Phi$ and $\Omega_R$ [21]. From Lemma 1, it appears that $\mu(\Omega_R, \Phi) < 1$ guarantees $\Phi\cap\Omega_R = \{\mathbf{0}_{L\times T}\}$. As it will become clear later on, tighter conditions on $\mu(\Omega_R, \Phi)$ will prove instrumental to guarantee exact recovery of $\{\mathbf{X}_0, \mathbf{A}_0\}$ by solving (P1).

To measure the incoherence among subsets of columns of $\mathbf{R}$, which is tightly related to the second condition in Lemma 1, the RICs come handy [15]. The constant $\delta_k(\mathbf{R})$ measures the extent to which a $k$-subset of columns of $\mathbf{R}$ behaves like an isometry. It is defined as the smallest value satisfying

$$c(1-\delta_k(\mathbf{R})) \leq \frac{\|\mathbf{Ru}\|^2}{\|\mathbf{u}\|^2} \leq c(1+\delta_k(\mathbf{R})) \tag{7}$$

for every $\mathbf{u} \in \mathbb{R}^F$ with $\|\mathbf{u}\|_0 \leq k$ and for some positive normalization constant $c < 1$ [15]. For later use, introduce $\theta_{s_1,s_2}(\mathbf{R})$ which measures "how orthogonal" are the subspaces generated by two disjoint column subsets of $\mathbf{R}$, with cardinality $s_1$ and $s_2$. Formally, $\theta_{s_1,s_2}(\mathbf{R})$ is the smallest value that satisfies

$$|\langle \mathbf{Ru}_1, \mathbf{Ru}_2\rangle| \leq c\theta_{s_1,s_2}(\mathbf{R})\|\mathbf{u}_1\|\|\mathbf{u}_2\| \tag{8}$$

for every $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^F$, where $\text{supp}(\mathbf{u}_1) \cap \text{supp}(\mathbf{u}_2) = \emptyset$ and $\|\mathbf{u}_1\|_0 \leq s_1, \|\mathbf{u}_2\|_0 \leq s_2$. The normalization constant $c$ plays the same role as in $\delta_k(\mathbf{R})$. A wide family of matrices with small RICs have been introduced in e.g., [15].

All the elements are now in place to state this section's main result.

*Proposition 1:* Assume that each column of $\mathbf{A}_0$ contains at most $k$ nonzero elements. If $\mu(\Omega_R, \Phi) < 1$ and $\delta_k(\mathbf{R}) < 1$, then $\Omega_R \cap \Phi = \{\mathbf{0}_{L\times T}\}$ and $\mathbf{RH} \neq \mathbf{0}_{L\times T}, \forall\, \mathbf{H} \in \Omega \backslash \{\mathbf{0}_{F\times T}\}$.

*Proof:* Suppose the intersection $\Omega_R \cap \Phi$ is nontrivial, meaning that it contains at least one nonzero matrix $\mathbf{H}_1 \in \mathbb{R}^{F\times T}$. Then $\mathcal{P}_\Phi(\mathbf{H}_1) = (\mathbf{H}_1)$, and consequently (6) gives rise to $\mu(\Omega_R, \Phi) = 1$ which is a contradiction. Likewise, suppose there exists a nonzero matrix $\mathbf{H}_2 \in \mathbb{R}^{F\times T}$ satisfying $\mathbf{RH}_2 = \mathbf{0}_{L\times T}$, and at least one of its columns contains $k > 0$ nonzero elements. Then, (7) leads to

$$c(1-\delta_k(\mathbf{R}))\|\mathbf{H}_2\|_F^2 \leq \|\mathbf{RH}_2\|_F^2 = 0 \leq c(1+\delta_k(\mathbf{R}))\|\mathbf{H}_2\|_F^2 \tag{9}$$

which implies $\delta_k(\mathbf{R}) \geq 1$, and contradicts the assumption $\delta_k(\mathbf{R}) < 1$. ■

## IV. Exact Recovery Via Convex Optimization

In addition to $\mu(\Omega_R, \Phi)$, there are other incoherence measures that play an important role in the conditions for exact recovery. Consider a feasible solution $\{\mathbf{X}_0 + a_{ij}\mathbf{R}\mathbf{e}_i\mathbf{e}_j', \mathbf{A}_0 - a_{ij}\mathbf{e}_i\mathbf{e}_j'\}$, where $(i,j) \notin \text{supp}(\mathbf{A}_0)$ and thus $a_{ij}\mathbf{e}_i\mathbf{e}_j' \notin \Omega$. It may then happen that $a_{ij}\mathbf{R}\mathbf{e}_i\mathbf{e}_j' \in \Phi$ and $\text{rank}(\mathbf{X}_0 + a_{ij}\mathbf{R}\mathbf{e}_i\mathbf{e}_j') = \text{rank}(\mathbf{X}_0) - 1$, while $\|\mathbf{A}_0 - a_{ij}\mathbf{e}_i\mathbf{e}_j'\|_0 = \|\mathbf{A}_0\|_0 + 1$, challenging identifiability when $\Phi$ and $\Omega_R$ are unknown. Similar complications will arise if $\mathbf{X}_0$ has a sparse row space that could be confused with the row space of $\mathbf{A}_0$. These issues motivate defining (recall $\mathbf{X}_0 = \mathbf{U\Sigma V}'$)

$$\gamma_R(\mathbf{U}) := \max_{i,j} \frac{\|\mathbf{P}_U\mathbf{R}\mathbf{e}_i\mathbf{e}_j'\|_F}{\|\mathbf{R}\mathbf{e}_i\mathbf{e}_j'\|_F}, \quad \gamma(\mathbf{V}) := \max_i \|\mathbf{P}_V\mathbf{e}_i\|_F$$

where $\gamma_R(\mathbf{U}), \gamma(\mathbf{V}) \leq 1$. The maximum of $\gamma_R(\mathbf{U})$ $[\gamma(\mathbf{V})]$ is attained when $\mathbf{R}\mathbf{e}_i\mathbf{e}_j'$ $[\mathbf{e}_i]$ is in the column [row] space of $\mathbf{X}_0$ for some $(i,j)$. Small values of $\gamma_R(\mathbf{U})$ and $\gamma(\mathbf{V})$ imply that the column and row spaces of $\mathbf{X}_0$ do not contain the columns of $\mathbf{R}$ and sparse vectors, respectively.

Another identifiability issue arises when $\mathbf{X}_0 = \mathbf{RH}$ for some sparse matrix $\mathbf{H} \in \Omega$. In this case, each column of $\mathbf{X}_0$ is spanned by a few columns of $\mathbf{R}$. Consider the parameter

$$\xi_R(\mathbf{U},\mathbf{V}) := \|\mathbf{R}'\mathbf{U}\mathbf{V}'\|_\infty = \max_{i,j} |\mathbf{e}_i{}'\mathbf{R}'\mathbf{U}\mathbf{V}\mathbf{e}_j|.$$

A small value of $\xi_R(\mathbf{U},\mathbf{V})$ implies that each column of $\mathbf{X}_0$ is spanned by sufficiently many columns of $\mathbf{R}$. To understand this property, consider the SVD $\mathbf{X}_0 = \mathbf{U\Sigma V}' = \sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i'$. The $k$th column of $\mathbf{X}_0$ is then $\sum_{i=1}^r \sigma_i\mathbf{u}_i v_{i,k}$, and its projection onto the $l$th column of $\mathbf{R}$ is

$$\left|\langle \mathbf{R}\mathbf{e}_l, \sum_{i=1}^r \sigma_i\mathbf{u}_i v_{i,k}\rangle\right| = \left|\sum_{i=1}^r \langle \mathbf{R}\mathbf{e}_l, \mathbf{u}_i\rangle \sigma_i v_{i,k}\right| \leq \sigma_{\max}\xi_R(\mathbf{U},\mathbf{V})$$

where $\sigma_{\max}$ is the largest singular value of $\mathbf{X}_0$. Since the energy of $\sum_{i=1}^r \sigma_i\mathbf{u}_i v_{i,k}$ is somehow allocated along the directions $\mathbf{R}\mathbf{e}_l$, if all the aforementioned projections can be made arbitrarily small, then sufficiently many nonzero terms in the expansion are needed to account for all this energy.

### A. Main Result

*Theorem 1:* Consider given matrices $\mathbf{Y} \in \mathbb{R}^{L\times T}$ and $\mathbf{R} \in \mathbb{R}^{L\times F}$ obeying $\mathbf{Y} = \mathbf{X}_0 + \mathbf{RA}_0 = \mathbf{U\Sigma V}' + \mathbf{RA}_0$, with $r := \text{rank}(\mathbf{X}_0)$ and $s := \|\mathbf{A}_0\|_0$. Assume that every row and column of $\mathbf{A}_0$ have at most $k$ nonzero elements, and that $\mathbf{R}$ has orthonormal rows. If the following conditions:

I) $\chi := \omega[(1-\mu(\Phi,\Omega_R))^2(1-\delta_k(\mathbf{R}))]^{-1} < 1/2$; and

II)

$$\lambda_{\max} := \sqrt{s^{-1}}\left[\alpha^{-1} - \mu(\Phi,\Omega_R)\sqrt{rc(1+\delta_k(\mathbf{R}))}\right] > \lambda_{\min} := \beta\xi_R(\mathbf{U},\mathbf{V})$$

hold, where

$$\begin{aligned}\omega &:= \theta_{1,1}(\mathbf{R})[\sqrt{2}k + s\gamma^2(\mathbf{V})] \\ &\quad + (1+\delta_1(\mathbf{R}))\left[\sqrt{2}k\gamma_R^2(\mathbf{U}) + k\gamma^2(\mathbf{V}) + s\gamma_R^2(\mathbf{U})\gamma^2(\mathbf{V})\right] \\ \alpha &:= 1 + \left(c^{-1}\omega^{-1}\chi - 1\right)^{1/2}, \quad \beta := (1-2\chi)^{-1}\end{aligned}$$

then there exists $\lambda \in (\lambda_{\min}, \lambda_{\max})$ for which the convex program (P1) exactly recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$.

Note that I) alone is already more stringent than the pair of conditions $\mu(\Omega_R,\Phi) < 1$ and $\delta_k(\mathbf{R}) < 1$ needed for local identifiability (cf., Proposition 1). Satisfaction of the conditions in Theorem 1 hinges upon the values of the incoherence parameters $\mu(\Omega_R,\Phi), \gamma_R(\mathbf{U}), \gamma(\mathbf{V}),\ \xi_R(\mathbf{U},\mathbf{V})$, and the RICs $\delta_k(\mathbf{R})$ and $\theta_{1,1}(\mathbf{R})$. In particular, $\{\omega,\alpha,\beta,\chi\}$ are increasing functions of these parameters, and it is readily observed from I) and II) that the smaller $\{\omega,\alpha,\beta\}$ are, the more likely the conditions are met. Furthermore, the incoherence parameters are increasing functions of the rank $r$ and sparsity level $s$. The RIC $\delta_k(\mathbf{R})$ is also an increasing function of $k$, the maximum number of nonzero elements per row/column of $\mathbf{A}_0$. Therefore, for sufficiently small values of $\{r,s,k\}$, the sufficient conditions of Theorem 1 can be indeed satisfied.

It is worth noting that not only $s$, but also the position of the nonzero entries in $\mathbf{A}_0$ plays an important role in satisfying I) and II). This is manifested through $k$, for which a small value indicates the entries of $\mathbf{A}_0$ are sufficiently spread out, i.e., most entries do not cluster along a few rows or columns of $\mathbf{A}_0$. Moreover, no restriction is placed on the magnitude of these entries, since as seen later on it is only the positions that affect optimal recovery via (P1).

*Remark 2 (Row Orthonormality of* $\mathbf{R}$*):* Assuming $\mathbf{RR}' = \mathbf{I}_L$ is equivalent to supposing that $\mathbf{R}$ is full-rank. This is because for a full row-rank $\mathbf{R} = \mathbf{U}_R\boldsymbol{\Sigma}_R\mathbf{V}_R{}'$, one can premultiply both sides of (1) with $\boldsymbol{\Sigma}_R^{-1}\mathbf{U}_R{}'$ to obtain $\tilde{\mathbf{R}} := \mathbf{V}_R{}'$ with orthonormal rows.

### B. Induced Recovery Results for Principal Components Pursuit and Compressed Sensing

Before delving into the proof of the main result, it is instructive to examine how the sufficient conditions in Theorem 1 simplify for the subsumed PCP and CS problems. In PCP, one has $\mathbf{R} = \mathbf{I}_L$, which implies $\Omega_R = \Omega$ and $\delta_k(\mathbf{R}) = \theta_{1,1}(\mathbf{R}) = 0$ so that one readily arrives at the following result.

*Corollary 1:* Consider given $\mathbf{Y} \in \mathbb{R}^{L\times T}$ obeying $\mathbf{Y} = \mathbf{X}_0 + \mathbf{A}_0 = \mathbf{U\Sigma V}' + \mathbf{A}_0$, with $r := \text{rank}(\mathbf{X}_0)$ and $s := \|\mathbf{A}_0\|_0$. If the following conditions:

$\mathcal{I}$) $\mu(\Phi,\Omega) + 2\sqrt{k(\gamma^2(\mathbf{U}) + \gamma^2(\mathbf{V}))} < 1$; and

$\mathcal{II}$) $\lambda_{\max} := \sqrt{s^{-1}}(\alpha^{-1} - \mu(\Phi,\Omega_R)\sqrt{r}) > \lambda_{\min} := \beta\xi(\mathbf{U},\mathbf{V})$

hold, where

$$\alpha := 1 + [(1-\mu(\Phi,\Omega))^{-2} - 1]^{1/2},$$
$$\beta := \left[1 - 4k(\gamma^2(\mathbf{U}) + \gamma^2(\mathbf{V}))(1-\mu(\Phi,\Omega))^{-2}\right]^{-1}$$

then there exists $\lambda \in (\lambda_{\min}, \lambda_{\max})$ for which the convex program (P1) with $\mathbf{R} = \mathbf{I}_L$ exactly recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$.

In Section VI, random matrices $\{\mathbf{X}_0, \mathbf{A}_0, \mathbf{R}\}$ drawn from natural ensembles are shown to satisfy I) and II) with high probability. In this case, it is possible to arrive at simpler conditions (depending only on $r$, $s$, and the matrix dimensions) for exact recovery in the context of PCP; see Remark 6 that compares Corollary 1 with the existing results for PCP. Corollary 1, on the other hand, offers general conditions stemming from a purely deterministic approach. The best deterministic recovery results for PCP appear to be those reported in [18].

In the CS setting, one has $\mathbf{X}_0 = \mathbf{0}_{L\times T}$, which implies $\mu(\Phi,\Omega_R) = \xi_R(\mathbf{U},\mathbf{V}) = \gamma_R(\mathbf{U}) = \gamma(\mathbf{V}) = 0$. As a result, Theorem 1 simply boils down to an RIC-dependent sufficient condition for the exact recovery of $\mathbf{A}_0$ as stated next.

*Corollary 2:* Consider given matrices $\mathbf{Y} \in \mathbb{R}^{L\times T}$ and $\mathbf{R} \in \mathbb{R}^{L\times F}$ obeying $\mathbf{Y} = \mathbf{R}\mathbf{A}_0$. Assume that the number of nonzero elements per column of $\mathbf{A}_0$ does not exceed $k$. If

$$\delta_k(\mathbf{R}) + k\theta_{1,1}(\mathbf{R}) < 1 \tag{10}$$

holds, then (P1) with $\mathbf{X} = \mathbf{0}_{L\times T}$ exactly recovers $\mathbf{A}_0$.

To place (10) in context, consider normalizing the rows of $\mathbf{R}$. For such a compression matrix, it is known that $\delta_k(\mathbf{R}) \leq (k-1)\theta_{1,1}(\mathbf{R})$; see, e.g., [43]. Using this bound together with (10), one arrives at the stricter condition $k < \frac{1}{2}\left(1 + \theta_{1,1}^{-1}(\mathbf{R})\right)$. This last condition is identical to the one reported in [23], which guarantees the success of $\ell_1$-norm minimization in recovering sparse solutions to under-determined systems of linear equations. The conditions have been improved in recent works; see, e.g., [43] and references therein.

## V. Proof of the Main Result

In what follows, conditions are first derived under which $\{\mathbf{X}_0, \mathbf{A}_0\}$ is the *unique* optimal solution of (P1). In essence, these conditions are expressed in terms of certain dual certificates. Then, Section V-B deals with the construction of a valid dual certificate.

### A. Unique Optimality Conditions

Recall the *nonsmooth* optimization problem (P1) and its Lagrangian

$$\mathcal{L}(\mathbf{X},\mathbf{A},\mathbf{M}) = \|\mathbf{X}\|_* + \lambda\|\mathbf{A}\|_1 + \langle \mathbf{M}, \mathbf{Y} - \mathbf{X} - \mathbf{RA}\rangle \tag{11}$$

where $\mathbf{M} \in \mathbb{R}^{L\times T}$ is the matrix of dual variables (multipliers) associated with the constraint in (P1). From the characterization of the subdifferential for nuclear- and $\ell_1$-norm (see, e.g., [10]), the subdifferential of the Lagrangian at $\{\mathbf{X}_0, \mathbf{A}_0\}$ is given by (recall that $\mathbf{X}_0 = \mathbf{U\Sigma V}'$)

$$\begin{aligned}\partial_{\mathbf{X}}\mathcal{L}(\mathbf{X}_0,\mathbf{A}_0,\mathbf{M}) = \{&\mathbf{UV}' + \mathbf{W} - \mathbf{M} : \\ &\|\mathbf{W}\| \leq 1, \quad \mathcal{P}_\Phi(\mathbf{W}) = \mathbf{0}_{L\times T}\} \quad (12)\\ \partial_{\mathbf{A}}\mathcal{L}(\mathbf{X}_0,\mathbf{A}_0,\mathbf{M}) = \{&\lambda\text{sign}(\mathbf{A}_0) + \lambda\mathbf{F} - \mathbf{R}'\mathbf{M} : \\ &\|\mathbf{F}\|_\infty \leq 1, \quad \mathcal{P}_\Omega(\mathbf{F}) = \mathbf{0}_{F\times T}\}. \quad (13)\end{aligned}$$

The optimality conditions for (P1) assert that $\{\mathbf{X}_0, \mathbf{A}_0\}$ is an optimal (not necessarily unique) solution if and only if

$$\mathbf{0}_{F\times T} \in \partial_{\mathbf{A}}\mathcal{L}(\mathbf{X}_0,\mathbf{A}_0,\mathbf{M}) \text{ and } \mathbf{0}_{L\times T} \in \partial_{\mathbf{X}}\mathcal{L}(\mathbf{X}_0,\mathbf{A}_0,\mathbf{M}).$$

This can be shown equivalent to finding the pair $\{\mathbf{W},\mathbf{F}\}$ that satisfies: i) $\|\mathbf{W}\| \leq 1$, $\mathcal{P}_\Phi(\mathbf{W}) = \mathbf{0}_{L\times T}$; ii) $\|\mathbf{F}\|_\infty \leq 1$, $\mathcal{P}_\Omega(\mathbf{F}) = \mathbf{0}_{F\times T}$; and iii) $\lambda\text{sign}(\mathbf{A}_0) + \lambda\mathbf{F} = \mathbf{R}'(\mathbf{UV}' + \mathbf{W})$. In general, i)–iii) may hold for multiple solution pairs. However, the next lemma asserts that a slight tightening of the optimality conditions i)–iii) leads to a *unique* optimal solution for (P1). See Appendix A for a proof.

*Lemma 2:* Assume that each column of $\mathbf{A}_0$ contains at most $k$ nonzero elements, as well as $\mu(\Omega_R,\Phi) < 1$ and $\delta_k(\mathbf{R}) < 1$. If there exists a dual certificate $\mathbf{\Gamma} \in \mathbb{R}^{L\times T}$ satisfying:

C1) $\mathcal{P}_\Phi(\mathbf{\Gamma}) = \mathbf{UV}'$

C2) $\mathcal{P}_\Omega(\mathbf{R}'\mathbf{\Gamma}) = \lambda\text{sgn}(\mathbf{A}_0)$

C3) $\|\mathcal{P}_{\Phi^\perp}(\mathbf{\Gamma})\| < 1$

C4) $\|\mathcal{P}_{\Omega^\perp}(\mathbf{R}'\mathbf{\Gamma})\|_\infty < \lambda$

then $\{\mathbf{X}_0, \mathbf{A}_0\}$ is the unique optimal solution of (P1).

The remainder of the proof deals with the construction of a dual certificate $\mathbf{\Gamma}$ that meets C1)–C4). To this end, tighter conditions [I) and II) in Theorem 1] for the existence of $\mathbf{\Gamma}$ are derived in terms of the incoherence parameters and the RICs. For the special case $\mathbf{R} = \mathbf{I}_L$, the conditions in Lemma 2 boil down to those in [17, Prop. 2] for PCP. However, the dual certificate construction techniques used in [17] do not carry over to the setting considered here, where a compression matrix $\mathbf{R}$ is present.

### B. Dual Certificate Construction

Condition C1) in Lemma 2 implies that $\mathbf{\Gamma} = \mathbf{UV}' + (\mathbf{I} - \mathbf{P}_U)\mathbf{X}(\mathbf{I} - \mathbf{P}_V)$, for arbitrary $\mathbf{X} \in \mathbb{R}^{L\times T}$ (cf., Remark 1). Upon defining $\mathbf{Z} := \mathbf{R}'(\mathbf{I} - \mathbf{P}_U)\mathbf{X}(\mathbf{I} - \mathbf{P}_U)$ and $\mathbf{B}_\Omega := \lambda\text{sign}(\mathbf{A}_0) - \mathcal{P}_\Omega(\mathbf{R}'\mathbf{UV}')$, C1) and C2) are equivalent to $\mathcal{P}_\Omega(\mathbf{Z}) = \mathbf{B}_\Omega$.

To express $\mathcal{P}_\Omega(\mathbf{Z}) = \mathbf{B}_\Omega$ in terms of the unrestricted matrix $\mathbf{X}$, first vectorize $\mathbf{Z}$ to obtain $\text{vec}(\mathbf{Z}) = [(\mathbf{I} - \mathbf{P}_V) \otimes \mathbf{R}'(\mathbf{I} - \mathbf{P}_U)]\,\text{vec}(\mathbf{X})$. Define $\mathbf{A} := (\mathbf{I} - \mathbf{P}_V) \otimes \mathbf{R}'(\mathbf{I} - \mathbf{P}_U)$ and an $s \times LT$ matrix $\mathbf{A}_\Omega$ formed with those $s$ rows of $\mathbf{A}$ associated with those elements in $\text{supp}(\mathbf{A}_0)$. Likewise, define $\mathbf{A}_{\Omega^\perp}$ which collects the remaining rows from

$\mathbf{A}$ such that $\mathbf{A} = \mathbf{\Pi}[\mathbf{A}'_{\Omega}, \mathbf{A}'_{\Omega^{\perp}}]'$ for a suitable row permutation matrix $\mathbf{\Pi}$. Finally, let $\mathbf{b}_{\Omega}$ be the vector of length $s$ containing those elements of $\mathbf{B}_{\Omega}$ with indices in $\text{supp}(\mathbf{A}_0)$. With these definitions, C1) and C2) can be expressed as $\mathbf{A}_{\Omega}\text{vec}(\mathbf{X}) = \mathbf{b}_{\Omega}$.

To upper-bound the left hand side of C3) in terms of $\mathbf{X}$, use the assumption $\mathbf{R}\mathbf{R}' = \mathbf{I}_L$ to arrive at

$$\begin{aligned}\|\mathcal{P}_{\Phi^{\perp}}(\mathbf{\Gamma})\| &= \|\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{X}(\mathbf{I}-\mathbf{P}_V)\|\\ &\leq \|\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{X}(\mathbf{I}-\mathbf{P}_V)\|_F = \|\mathbf{A}\text{vec}(\mathbf{X})\|.\end{aligned}$$

Similarly, the left hand side of C4) can be bounded as

$$\begin{aligned}\|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{\Gamma})\|_{\infty} &= \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{Z}) + \mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}\\ &\leq \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{Z})\|_{\infty} + \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}\\ &= \|\mathbf{A}_{\Omega^{\perp}}\text{vec}(\mathbf{X})\|_{\infty} + \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}.\end{aligned}$$

In a nutshell, if one can find $\mathbf{X} \in \mathbb{R}^{L\times T}$ such that:

c1) $\mathbf{A}_{\Omega}\text{vec}(\mathbf{X}) = \mathbf{b}_{\Omega}$

c2) $\|\mathbf{A}\text{vec}(\mathbf{X})\| < 1$

c3) $\|\mathbf{A}_{\Omega^{\perp}}\text{vec}(\mathbf{X})\|_{\infty} + \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty} < \lambda$

hold for some positive $\lambda$, then C1)–C4) would be satisfied as well.

The final steps of the proof entail: i) finding an appropriate candidate solution $\hat{\mathbf{X}}$ such that c1) holds; and ii) deriving conditions in terms of the incoherence parameters and RICs that guarantee $\hat{\mathbf{X}}$ meets the required bounds in c2) and c3) for a range of $\lambda$ values. The following lemma is instrumental to accomplishing i), and its proof can be found in Appendix B.

*Lemma 3:* Assume that each column of $\mathbf{A}_0$ contains at most $k$ nonzero elements, as well as $\mu(\Omega_R, \Phi) < 1$ and $\delta_k(\mathbf{R}) < 1$. Then, matrix $\mathbf{A}_{\Omega}$ has full row rank, and its minimum singular value is bounded below as

$$\sigma_{\min}(\mathbf{A}'_{\Omega}) \geq c^{1/2}(1-\delta_k(\mathbf{R}))^{1/2}(1-\mu(\Phi,\Omega_R)).$$

According to Lemma 3, the least-norm (LN) solution $\hat{\mathbf{X}}_{\text{LN}} := \arg\min_{\mathbf{X}}\left\{\|\mathbf{X}\|_F^2 : \mathbf{A}_{\Omega}\text{vec}(\mathbf{X}) = \mathbf{b}_{\Omega}\right\}$ exists and is given by

$$\text{vec}(\hat{\mathbf{X}}_{\text{LN}}) = \mathbf{A}'_{\Omega}\left(\mathbf{A}_{\Omega}\mathbf{A}'_{\Omega}\right)^{-1}\mathbf{b}_{\Omega}. \quad (14)$$

*Remark 3 (Candidate Dual Certificate):* From the arguments at the beginning of this section, the candidate dual certificate is $\hat{\mathbf{\Gamma}} := \mathbf{U}\mathbf{V}' + (\mathbf{I}-\mathbf{P}_U)\hat{\mathbf{X}}_{\text{LN}}(\mathbf{I}-\mathbf{P}_V)$.

The LN solution is an attractive choice, since it facilitates satisfying c2) and c3) which require norms of $\text{vec}(\mathbf{X})$ to be small. Substituting the LN solution (14) into the left hand side of c2) yields (define $\mathbf{Q} := \mathbf{A}_{\Omega^{\perp}}\mathbf{A}'_{\Omega}\left(\mathbf{A}_{\Omega}\mathbf{A}'_{\Omega}\right)^{-1}$ for notational brevity)

$$\begin{aligned}\|\mathbf{A}\text{vec}(\hat{\mathbf{X}}_{\text{LN}})\| &= \left\|\begin{pmatrix}\mathbf{A}_{\Omega}\\ \mathbf{A}_{\Omega^{\perp}}\end{pmatrix}\mathbf{A}'_{\Omega}\left(\mathbf{A}_{\Omega}\mathbf{A}'_{\Omega}\right)^{-1}\mathbf{b}_{\Omega}\right\|\\ &= \left\|\begin{pmatrix}\mathbf{I}\\ \mathbf{Q}\end{pmatrix}\mathbf{b}_{\Omega}\right\| \leq (1+\|\mathbf{Q}\|)\,\|\mathbf{b}_{\Omega}\|. \quad (15)\end{aligned}$$

Moreover, substituting (14) into the left hand side of c3) results in

$$\begin{aligned}\|\mathbf{Q}\mathbf{b}_{\Omega}\|_{\infty} + \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty} &\leq \|\mathbf{Q}\|_{\infty,\infty}\|\mathbf{b}_{\Omega}\|_{\infty}\\ &\quad + \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}. \quad (16)\end{aligned}$$

Next, upper-bounds are obtained for $\|\mathbf{Q}\|$ and $\|\mathbf{Q}\|_{\infty,\infty}$; see Appendix C for a proof.

*Lemma 4:* Assume that each column and row of $\mathbf{A}_0$ contains at most $k$ nonzero elements. If $\mu(\Omega_R, \Phi) < 1$ and $\delta_k(\mathbf{R}) < 1$ hold, then

$$\|\mathbf{Q}\| \leq \nu_1 := \left[\frac{1}{c(1-\delta_k(\mathbf{R}))(1-\mu(\Omega_R,\Phi))^2} - 1\right]^{1/2}.$$

If the tighter condition I) holds instead, then

$$\|\mathbf{Q}\|_{\infty,\infty} \leq \nu_2 := \frac{\omega}{(1-\mu(\Omega_R,\Phi))^2(1-\delta_k(\mathbf{R})) - \omega}.$$

Going back to (15)–(16), note that $\|\mathbf{B}_{\Omega}\|_{\infty} = \|\mathbf{b}_{\Omega}\|_{\infty}$ and $\|\mathbf{B}_{\Omega}\|_F = \|\mathbf{b}_{\Omega}\|$, which can be, respectively, upper-bounded as

$$\begin{aligned}\|\mathbf{B}_{\Omega}\|_{\infty} &= \|\lambda\text{sign}(\mathbf{A}_0) - \mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}\\ &\leq \lambda + \|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty} \quad (17)\\ \|\mathbf{B}_{\Omega}\|_F &= \|\lambda\text{sign}(\mathbf{A}_0) - \mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F\\ &\leq \lambda\sqrt{s} + \|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F. \quad (18)\end{aligned}$$

Finally, $\|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F$ itself can be bounded above as

$$\begin{aligned}\|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F^2 &= |\langle\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}'), \mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\rangle|\\ &\overset{(a)}{=} |\langle\mathbf{R}'\mathbf{U}\mathbf{V}', \mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\rangle|\\ &= |\langle\mathbf{U}\mathbf{V}', \mathbf{R}\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\rangle|\\ &\overset{(b)}{=} |\langle\mathcal{P}_{\Phi}(\mathbf{U}\mathbf{V}'), \mathcal{P}_{\Phi}(\mathbf{R}\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}'))\rangle|\\ &\overset{(c)}{\leq} \|\mathcal{P}_{\Phi}(\mathbf{U}\mathbf{V}')\|_F\|\mathcal{P}_{\Phi}(\mathbf{R}\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}'))\|_F\\ &\overset{(d)}{\leq} \|\mathbf{U}\mathbf{V}'\|_F\,\mu(\Phi,\Omega_R)\|\mathbf{R}\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F\\ &\overset{(e)}{\leq} \sqrt{r}\mu(\Phi,\Omega_R)c^{1/2}(1+\delta_k(\mathbf{R}))^{1/2}\\ &\quad\times\|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F \quad (19)\end{aligned}$$

where $(a)$ is due to (5), $(b)$ follows because $\mathbf{U}\mathbf{V}' \in \Phi$ (thus $\mathcal{P}_{\Phi}(\mathbf{U}\mathbf{V}') = \mathbf{U}\mathbf{V}'$) and from the property in (5). Moreover, $(c)$ is a direct result of the Cauchy–Schwarz inequality, while $(d)$ and $(e)$ come from (6) and (7), respectively, and the assumption that number of nonzero elements per column of $\mathbf{A}_0$ does not exceed $k$. All in all, $\|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_F \leq \sqrt{r}\mu(\Phi,\Omega_R)c^{1/2}(1+\delta_k(\mathbf{R}))^{1/2}$ and (18) becomes

$$\|\mathbf{B}_{\Omega}\|_F \leq \lambda\sqrt{s} + \sqrt{r}\mu(\Phi,\Omega_R)c^{1/2}(1+\delta_k(\mathbf{R}))^{1/2}. \quad (20)$$

Upon substituting (17), (20) and the bounds in Lemma 4 into (15) and (16), one finds that c2) and c3) hold if there exists $\lambda > 0$ such that

$$(1+\nu_1)\left[\lambda\sqrt{s} + \sqrt{r}\mu(\Omega_R,\Phi)c^{1/2}(1+\delta_k(\mathbf{R}))^{1/2}\right] < 1 \quad (21a)$$

$$\nu_2\left(\lambda + \|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}\right) + \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty} < \lambda \quad (21b)$$

hold. Recognizing that $\xi_R(\mathbf{U},\mathbf{V}) = \max\{\|\mathcal{P}_{\Omega}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}, \|\mathcal{P}_{\Omega^{\perp}}(\mathbf{R}'\mathbf{U}\mathbf{V}')\|_{\infty}\}$the left hand side of (21b) can be further bounded. After straightforward manipulations, one deduces that conditions (21a) and (21b) are satisfied for $\lambda \in (\lambda_{\min}, \lambda_{\max})$ if $\nu_2 < 1$, where

$$\lambda_{\min} := \left(\frac{1+\nu_2}{1-\nu_2}\right)\xi_R(\mathbf{U},\mathbf{V})$$

$$\lambda_{\max} := \frac{1}{\sqrt{s}}\left[(1+\nu_1)^{-1} - \sqrt{r}\mu(\Omega_R,\Phi)c^{1/2}(1+\delta_k(\mathbf{R}))^{1/2}\right].$$

Clearly, it is still necessary to ensure $\lambda_{\max} > \lambda_{\min}$ so that the LN solution (14) meets the requirements c1)–c3) [equivalently, $\hat{\mathbf{\Gamma}}$ in Remark 3 satisfies C1)–C4) from Lemma 2]. Condition $\lambda_{\max} > \lambda_{\min}$ is equivalent to II) in Theorem 1, and the proof is now complete.

*Remark 4 (Satisfiability):* From a high-level vantage point, Theorem 1 asserts that (P1) recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$ when the components $\mathbf{X}_0$ and $\mathbf{R}\mathbf{A}_0$ are sufficiently incoherent, and the compression matrix $\mathbf{R}$ has good restricted isometry properties. It should be noted though that given a triplet $\{\mathbf{X}_0, \mathbf{A}_0, \mathbf{R}\}$ in general, one cannot directly check whether the sufficient conditions I) and II) hold, since, e.g., $\delta_k(\mathbf{R})$ is NP-hard to compute [15]. This motivates finding a class of (possibly random) matrices $\{\mathbf{X}_0, \mathbf{A}_0, \mathbf{R}\}$ satisfying I) and II), the subject dealt with next.

## VI. Matrices Satisfying the Conditions for Exact Recovery

This section investigates triplets $\{\mathbf{X}_0, \mathbf{A}_0, \mathbf{R}\}$ satisfying the conditions of Theorem 1, henceforth termed admissible matrices. Specifically, it will be shown that low-rank, sparse, and compression matrices drawn from certain random ensembles satisfy the sufficient conditions of Theorem 1 with high probability.

### *A. Uniform Sparsity Model*

Matrix $\mathbf{A}_0$ is said to be generated according to the *uniform sparsity* model, when drawn uniformly at random from the collection of all matrices with support size $s$. There is no restriction on the amplitude of the nonzero entries. An attractive property of this model is that it guarantees (with high probability) that no single row or column will monopolize most nonzero entries of $\mathbf{A}_0$, for sufficiently large $\mathbf{A}_0$ and appropriate scaling of the sparsity level. This property is formalized in the following lemma (for simplicity in exposition, it is henceforth assumed that $\mathbf{A}_0$ is a square matrix, i.e., $F = T$).

*Lemma 5 [17]:* If $\mathbf{A}_0 \in \mathbb{R}^{F \times F}$ is generated according to the uniform sparsity model with $\|\mathbf{A}_0\|_0 = s$, then the maximum number $k$ of nonzero elements per column or row of $\mathbf{A}_0$ is bounded as

$$k \le \frac{s}{F}\log(F)$$

with probability higher than $1 - \mathcal{O}(F^{-\zeta})$, for $s = \mathcal{O}(\zeta F)$.

In practice, it is simpler to work with the Bernoulli model that specifies $\mathrm{supp}(\mathbf{A}_0) = \{(i,j) : b_{i,j} = 1\}$, where $\{b_{i,j}\}$ are independent and identically distributed (i.i.d.) Bernoulli random variables taking value one with probability $\pi := s/F^2$, and zero with probability $1 - \pi$. There are three important observations regarding the Bernoulli model. First, $|\mathrm{supp}(\mathbf{A}_0)|$ is a random variable, whose expected value is $s$ and matches the uniform sparsity model. Second, arguing as in [11, Lemma 2.2], one can claim that if (P1) exactly recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$ from data $\mathbf{Y} = \mathbf{X}_0 + \mathbf{R}\mathbf{A}_0$, it will also exactly recover $\{\mathbf{X}_0, \check{\mathbf{A}}_0\}$ from $\check{\mathbf{Y}} = \mathbf{X}_0 + \mathbf{R}\check{\mathbf{A}}_0$ when $\mathrm{supp}(\check{\mathbf{A}}_0) \subseteq \mathrm{supp}(\mathbf{A}_0)$ and the nonzero entries coincide. Third, following the logic of [14, Sec. II.C], one can prove that the failure rate[1] for the uniform sparsity model is bounded by twice the failure rate corresponding to the Bernoulli model. As a result, any recovery guarantee established for the Bernoulli model holds for the uniform sparsity model as well.

In addition to the bound for $k$ in Lemma 5, the Bernoulli model can be used to bound $\mu(\Phi, \Omega_R)$ in terms of the incoherence parameters $\{\gamma_R(\mathbf{U}), \gamma(\mathbf{V})\}$ and the RIC $\delta_k(\mathbf{R})$. For a proof, see Appendix D.

*Lemma 6:* Let $\Lambda := \sqrt{c(1+\delta_1(\mathbf{R}))}\left[\gamma_R^2(\mathbf{U}) + \gamma^2(\mathbf{V})\right]^{1/2}$ and $n := \max\{L, F\}$. Suppose $\mathbf{A}_0 \in \mathbb{R}^{F \times F}$ is generated according to the Bernoulli model with $\Pr(b_{i,j} = 1) = \pi$, and $\mathbf{R}\mathbf{R}' = \mathbf{I}_L$. Then, there exist positive constants $C$ and $\tau$ such that

$$\mu(\Phi, \Omega_R) \le \sqrt{c^{-1}(1-\delta_k(\mathbf{R}))^{-1}\pi} \times \left[C\Lambda\sqrt{\log(LF)/\pi} + \tau\Lambda\log(n) + 1\right]^{1/2} \tag{22}$$

holds with probability at least $1 - n^{-C\pi\Lambda\tau}$ if $\delta_k(\mathbf{R})$ and the right-hand side of (22) do not exceed one.[2]

Consider (22) when $\Lambda$ is small enough so that the quantity inside the square brackets is close to one. One obtains $\mu(\Phi, \Omega_R) \le \sqrt{c^{-1}(1-\delta_k(\mathbf{R}))^{-1}\pi}$, which reduces to the bound $\mu(\Phi, \Omega) \le \sqrt{\pi}$ derived in [11, Sec. 2.5] for the special case $\mathbf{R} = \mathbf{I}_L$. Hence, the price paid in terms of coherence increase due to $\mathbf{R}$ is roughly $\sqrt{c^{-1}(1-\delta_k(\mathbf{R}))^{-1}} > 1$. As expected, (22) also shows that for $\mathbf{R}$ with small RICs the incoherence between subspaces $\Phi$ and $\Omega_R$ becomes smaller, and identifiability is more likely.

The result in Lemma 6 allows one to "eliminate" $\mu(\Phi, \Omega_R)$ from the sufficient conditions in Theorem 1, which can thus be expressed only in terms of $\{\gamma_R(\mathbf{U}), \gamma(\mathbf{V}), \xi_R(\mathbf{U}, \mathbf{V})\}$ and the RICs of $\mathbf{R}$. In the following sections, random low-rank and compression matrices giving rise to small incoherence parameters and RICs are described.

### *B. Random Orthogonal Model*

Among other implications, matrices $\mathbf{X}_0$ and $\mathbf{R}$ with small $\gamma_R(\mathbf{U})$ and $\xi_R(\mathbf{U}, \mathbf{V})$ are such that the columns of $\mathbf{R}$ (approximately) fall outside the column space of $\mathbf{X}_0$. From a design perspective, this suggests that the choice of an admissible $\mathbf{X}_0$ (or, in general, an ensemble of low-rank matrices) should take into account the structure of $\mathbf{R}$, and vice versa. However, in the interest of simplicity, one could seek conditions dealing with $\mathbf{X}_0$ and $\mathbf{R}$ *separately*, that still ensure $\gamma_R(\mathbf{U})$ and $\xi_R(\mathbf{U}, \mathbf{V})$ are small. This way one can benefit from the existing theory on incoherent low-rank matrices developed in the context of matrix completion [13], and matrices with small RICs useful for CS [14], [43]. Admittedly, the price paid is in terms of stricter conditions that will reduce the set of admissible matrices.

[1] The failure rate is defined as $\Pr(\hat{\mathbf{A}} \neq \mathbf{A}_0)$, where $\hat{\mathbf{A}}$ is the solution of (P1).

[2] Even though one has $n = F$ and $\pi = s/F^2$ in the problem studied here, Lemma 6 is stated using $n$ and $\pi$ to retain generality.

In this direction, the next lemma bounds $\gamma_R(\mathbf{U})$ and $\xi_R(\mathbf{U},\mathbf{V})$ in terms of $\gamma(\mathbf{U}) := \max_i \|\mathbf{P}_U\mathbf{e}_i\|$, $\gamma(\mathbf{V})$ and $\delta_k(\mathbf{R})$.

*Lemma 7:* If $\eta(\mathbf{R}) := \max_i \|\mathbf{R}\mathbf{e}_i\|_1/\|\mathbf{R}\mathbf{e}_i\|$, it then holds that

$$\gamma_R(\mathbf{U}) \leq \eta(\mathbf{R})\gamma(\mathbf{U}) \tag{23}$$

$$\xi_R(\mathbf{U},\mathbf{V}) \leq \sqrt{c(1+\delta_1(\mathbf{R}))}\eta(\mathbf{R})\gamma(\mathbf{U})\gamma(\mathbf{V}). \tag{24}$$

*Proof:* Starting from the definition

$$\begin{aligned}\gamma_R(\mathbf{U}) &= \max_i \frac{\|\mathbf{P}_U\mathbf{R}\mathbf{e}_i\|}{\|\mathbf{R}\mathbf{e}_i\|} = \max_i \frac{\|\mathbf{P}_U\sum_\ell \mathbf{e}_\ell\mathbf{e}_\ell'\mathbf{R}\mathbf{e}_i\|}{\|\mathbf{R}\mathbf{e}_i\|}\\ &\overset{(a)}{\leq} \max_i \frac{\sum_\ell \|\mathbf{P}_U\mathbf{e}_\ell\||\mathbf{e}_\ell'\mathbf{R}\mathbf{e}_i|}{\|\mathbf{R}\mathbf{e}_i\|} \overset{(b)}{\leq} \gamma(\mathbf{U})\max_i \frac{\|\mathbf{R}\mathbf{e}_i\|_1}{\|\mathbf{R}\mathbf{e}_i\|}\end{aligned} \tag{25}$$

where $(\mathrm{a})$ follows from the Cauchy–Schwarz inequality, and $(\mathrm{b})$ from the definition of $\gamma(\mathbf{U})$.

Likewise, applying the definition of $\xi_R(\mathbf{U},\mathbf{V})$, one obtains

$$\begin{aligned}\xi_R(\mathbf{U},\mathbf{V}) &= \max_{i,j} |\mathbf{e}_i'\mathbf{R}'\mathbf{U}\mathbf{V}'\mathbf{e}_j|\\ &\overset{(c)}{\leq} \max_i \|\mathbf{U}'\mathbf{R}\mathbf{e}_i\| \max_i \|\mathbf{V}'\mathbf{e}_j\|\\ &\leq \sqrt{c(1+\delta_1(\mathbf{R}))}\gamma_R(\mathbf{U})\gamma(\mathbf{V})\\ &\overset{(d)}{\leq} \sqrt{c(1+\delta_1(\mathbf{R}))}\eta(\mathbf{R})\gamma(\mathbf{U})\gamma(\mathbf{V})\end{aligned} \tag{26}$$

where $(\mathrm{c})$ follows from the Cauchy–Schwarz inequality, and $(\mathrm{d})$ is due to (25). ■

The bounds (23) and (24) are proportional to $\gamma(\mathbf{U})$ and $\gamma(\mathbf{V})$. This prompts one to consider incoherent rank-$r$ matrices $\mathbf{X}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ generated from the *random orthogonal* model, which is specified as follows. The singular vectors forming the columns of $\mathbf{U}$ and $\mathbf{V}$ are drawn uniformly at random from the collection of rank-$r$ partial isometries in $\mathbb{R}^{L\times r}$ and $\mathbb{R}^{F\times r}$, respectively. There is no need for $\mathbf{U}$ and $\mathbf{V}$ to be statistically independent, and no restriction in placed on the singular values in the diagonal of $\mathbf{\Sigma}$. The adequacy of the random orthogonal model in generating incoherent low-rank matrices is justified by the following lemma (recall $T = F \geq L$).

*Lemma 8 [17]:* If $\mathbf{X}_0 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \in \mathbb{R}^{L\times F}$ is generated according to the random orthogonal model with $\mathrm{rank}(\mathbf{X}_0) = r$, then

$$\max\{\gamma(\mathbf{U}),\gamma(\mathbf{V})\} \leq \sqrt{\frac{\max\{r,\log(F)\}}{F}}$$

with probability exceeding $1 - \mathcal{O}(F^{-3}\log(F))$.

### C. Random Compressive Matrices

With reference to Lemma 7 [cf., (23) and (24)], it is clear that an incoherent $\mathbf{X}_0$ alone may not suffice to yield small $\gamma_R(\mathbf{U})$ and $\xi_R(\mathbf{U},\mathbf{V})$. In addition, $\eta(\mathbf{R}) \in [1,\sqrt{L}]$ should be as close as possible to one. This can be achieved, e.g., when $\mathbf{R}$ is sparse across each column. Note that the lower bound of unity is attained when $\mathbf{R}$ has at most a single nonzero element per column, as it is the case when $\mathbf{R} = \mathbf{I}_L$.

The aforementioned observations motivate considering block-diagonal compression matrices $\mathbf{R} \in \mathbb{R}^{L\times F}$, consisting of blocks $\{\mathbf{R}_i \in \mathbb{R}^{\ell\times f}\}$ where $\ell \leq f$. The number of blocks is $n_b := F/f$ assuming that $f$ divides $F$. The $i$th block is generated according to the *bounded orthonormal* model as follows; see, e.g., [43]. For some positive constant $K$, (deterministically) choose a unitary matrix $\mathbf{\Psi} \in \mathbb{R}^{f\times f}$ with bounded entries

$$\max_{(t,k)\in\mathcal{F}\times\mathcal{F}} |\mathbf{\Psi}_{t,k}| \leq K \tag{27}$$

where $\mathcal{F} := \{1,\ldots,f\}$. For each $i = 1,\ldots,n_b$ form $\mathbf{R}_i := \mathbf{\Theta}_{T^{(i)}}\mathbf{\Psi}$, where $\mathbf{\Theta}_{T^{(i)}} := [\mathbf{e}_{t_1^{(i)}},\ldots,\mathbf{e}_{t_\ell^{(i)}}]' \in \mathbb{R}^{\ell\times f}$ is a random row subsampling matrix that selects the rows of $\mathbf{\Psi}$ indexed by $\mathcal{T}^{(i)} := \{t_1^{(i)},\ldots,t_\ell^{(i)}\} \subset \mathcal{F}$. In words, $\mathbf{\Theta}_{T^{(i)}}$ is formed by those $\ell$ rows of $\mathbf{I}_f$ indexed by $\mathcal{T}^{(i)}$. The row indices in $\mathcal{T}^{(i)}$ are selected independently at random, with uniform probability $1/f$ from $\mathcal{F}$. By construction, $\mathbf{R}_i\mathbf{R}_i' = \mathbf{I}_\ell$, $i = 1,\ldots,n_b$, which ensures $\mathbf{R}\mathbf{R}' = \mathbf{I}_L$ as required by Theorem 1. Most importantly, the next lemma states that such a construction of $\mathbf{R}_i$ leads to small RICs with high probability; see, e.g., [43] for the proof.

*Lemma 9 [43]:* Let $\mathbf{R}_i \in \mathbb{R}^{\ell\times f}$ be generated according to the bounded orthonormal model. If for some $k_i \in [1,f]$, $\epsilon \in (0,1)$ and $\mu \in (0,1/2]$, the following condition

$$\frac{\ell}{\log(10\ell)} \geq DK^2\mu^{-2}s\log^2(100k_i)\log(4f)\log(7\epsilon^{-1}) \tag{28}$$

holds where the constant $D \leq 243,150$, then $\delta_{k_i}(\mathbf{R}_i) \leq \mu$ with probability greater than $1 - \epsilon$.

Lemma 9 asserts that for large enough $\ell$, the RIC $\delta_{k_i}(\mathbf{R}_i) = \mathcal{O}(\log(100k_i)\log(10\ell)\log(4f)^{1/2}\sqrt{k_i/\ell})$ with overwhelming probability.

Let $k_i$ denote the maximum number of nonzero elements per "trimmed" column of $\mathbf{A}_0$, the trimming being defined by the block of rows of $\mathbf{A}_0$ that are multiplied by $\mathbf{R}_i$ when carrying out the product $\mathbf{R}\mathbf{A}_0$. With these definitions, the RIC of $\mathbf{R}$ is bounded as $\delta_k(\mathbf{R}) \leq \max_i\{\delta_{k_i}(\mathbf{R}_i)\}$. For $\delta_k(\mathbf{R})$ to be small as required by Theorem 1, $k_i$ should be much smaller than $\ell$. Since $\mathbf{A}_0$ is generated according to the uniform sparsity model outlined in Section VI-A, its nonzero elements are uniformly spread across rows and columns as per Lemma 5. Formally, it holds that $k_i \leq \kappa := (s/Fn_b)\log(Fn_b)$ with probability $1 - \mathcal{O}([Fn_b]^{-\zeta})$, where $s = \|\mathbf{A}_0\|_0 = \zeta Fn_b$; see, e.g., [8]. Accordingly, from Lemma 9, one can infer that $\delta_k(\mathbf{R}) = \mathcal{O}(\log(100\kappa)\log(10\ell)\log(4f)^{1/2}\sqrt{\kappa/\ell})$ with high probability. Note that the bound for $\delta_k(\mathbf{R})$ depends on $k$ through the variable $s$ in $\kappa$, and the relationship between $s$ and $k$ in Lemma 5. Regarding the RIC $\theta_{1,1}(\mathbf{R})$, it is bounded as $\theta_{1,1}(\mathbf{R}) \leq \delta_2(\mathbf{R})$ [15]. The normalization constant $c$ in (7) and (8) also equals $L/F \ll 1$. Recalling $\eta(\mathbf{R})$ (cf., Lemma 7) which was subject of the initial discussion in this section, it turns out that for such a construction of $\mathbf{R}$, one obtains $\eta(\mathbf{R}) \leq \sqrt{\ell} \ll \sqrt{L}$.

*Remark 5 (Row and Column Permutations):* The class of admissible compression matrices can be extended to matrices which are block diagonal up to row and column permutations. Let $\mathbf{\Pi}_r$ ($\mathbf{\Pi}_c$) denote, respectively, the row (column) permutation matrices that render $\mathbf{R}$ block diagonal. Instead of (1), consider $\mathbf{\Pi}_r\mathbf{Y} = \mathbf{\Pi}_r\mathbf{X}_0 + \mathbf{\Pi}_r\mathbf{R}\mathbf{\Pi}_c\mathbf{\Pi}_c'\mathbf{A}_0$ and note that $\mathbf{\Pi}_r\mathbf{X}_0$ has the same coherence parameters as $\mathbf{X}_0$, while $\mathbf{\Pi}_r \mathbf{R} \mathbf{\Pi}_c$ has the same RICs as $\mathbf{R}$, and $\mathbf{\Pi}_c' \mathbf{A}_0$ is still uniformly sparse. Thus, one can feed the transformed data to (P1), and since $\mathbf{\Pi}_r$ and $\mathbf{\Pi}_c$ are invertible, $\{\mathbf{X}_0, \mathbf{A}_0\}$ can be readily obtained from the recovered $\{\mathbf{\Pi}_r \mathbf{X}_0, \mathbf{\Pi}_c' \mathbf{A}_0\}$.

### *D. Closing the Loop*

According to Lemmata 6 and 7, the incoherence parameters $\mu(\Phi, \Omega_R)$, $\gamma_R(\mathbf{U})$, and $\xi_R(\mathbf{U}, \mathbf{V})$ which play a critcal role toward exact decomposability in Theorem 1, can be upper-bounded in terms of $\gamma(\mathbf{U})$ and $\gamma(\mathbf{V})$. For random matrices $\{\mathbf{X}_0, \mathbf{A}_0, \mathbf{R}\}$ drawn from specific ensembles, Lemmata 5, 8, and 9 assert that the incoherence parameters $\gamma(\mathbf{U})$ and $\gamma(\mathbf{V})$ as well as the RICs $\delta_k(\mathbf{R})$ and $\theta_{1,1}(\mathbf{R})$, are bounded above in terms of $r = \mathrm{rank}(\mathbf{X}_0)$, the degree of sparsity $s = \|\mathbf{A}_0\|_0$, and the underlying matrix dimensions $L, F, \ell, f$. Alternative sufficient conditions for exact recovery, expressible only in terms of the aforementioned basic parameters, can be obtained by combining the bounds of this section along with I) and II) in Theorem 1 . Hence, in order to guarantee that (P1) recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$ with high probability and for given matrix dimensions, it suffices to check feasibility of a set of inequalities in $r$ and $s$.

To this end, focus on the asymptotic case where $L$ and $F$ are large enough, while $F = T$ for simplicity in exposition. Recall the conditions of Theorem 1 and suppose $\delta_k(\mathbf{R}) = o(1)$ and $\mu(\Phi, \Omega_R) = o(1)$. This results in $\alpha \approx \sqrt{F/L}$ and $\chi \approx \omega$ when $L \ll F$. Satisfaction of I) and II) then requires $\mathcal{O}(1)$ summands in both sides of II) when multiplied with $\alpha\sqrt{s}$, which gives rise to $\xi_R(\mathbf{U}, \mathbf{V}) = \mathcal{O}(\sqrt{L/Fs})$, $\mu(\Phi, \Omega_R) = \mathcal{O}(1/\sqrt{r})$, and $\omega = \mathcal{O}(1) < 1$. The latter which is indeed the bottleneck constraint can be satisfied if $\theta_{1,1}(\mathbf{R}) = \mathcal{O}(1/k)$, $\theta_{1,1}(\mathbf{R})\gamma^2(\mathbf{V}) = \mathcal{O}(1/s)$, $\gamma_R^2(\mathbf{U}) = \mathcal{O}(1/k)$, $\gamma^2(\mathbf{V}) = \mathcal{O}(1/k)$, and $\gamma_R^2(\mathbf{U})\gamma_R^2(\mathbf{V}) = \mathcal{O}(1/s)$. Utilizing the bounds in Lemmata 6–9 establishes the next corollary.

*Corollary 3:* Consider given matrices $\mathbf{Y} \in \mathbb{R}^{L \times F}$ and $\mathbf{R} \in \mathbb{R}^{L \times F}$ obeying $\mathbf{Y} = \mathbf{X}_0 + \mathbf{R}\mathbf{A}_0$, where $r := \mathrm{rank}(\mathbf{X}_0)$ and $s := \|\mathbf{A}_0\|_0$. Suppose that: (i) $\mathbf{X}_0$ is generated according to the random orthogonal model; (ii) $\mathbf{A}_0$ is generated according to the uniform sparsity model; and (iii) $\mathbf{R} = \mathrm{bdiag}(\mathbf{R}_1, \ldots, \mathbf{R}_{n_b})$ with blocks $\mathbf{R}_i \in \mathbb{R}^{\ell \times f}$ generated according to the bounded orthogonal model. Define $\tilde{r} := \max\{r, \log(F)\}$. If $r$ and $s$ satisfy:

i) $\tilde{r} \precsim \frac{F}{\ell}$

ii) $s \precsim \min\left\{\frac{F^2}{\ell \log(F)\tilde{r}}, \frac{F^2}{\tilde{r}^2}, \frac{F\sqrt{\ell}}{\log(10\ell)\log^{1/2}(4f)\tilde{r}}\right\}$

iii) $\sqrt{s}\log\left(100\frac{sf}{F^2}\log\left(\frac{F^2}{f}\right)\right) \prec \left[\frac{F^2\ell}{f\log(F^2/f)\log^2(f)}\right]^{1/2}$

there is a positive $\lambda$ for which (P1) recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$ with high probability.

*Remark 6 (Results for PCP):* For an ensemble of random matrices $\{\mathbf{X}_0, \mathbf{A}_0\}$, the induced recovery results for PCP in Corollary 1 are simplified and compared here with those obtained in [11], [17], and [18]. To be aligned with [11] and [18], the $\rho$-incoherent low-rank matrix model in [11] is adopted for $\mathbf{X}_0$, where $\gamma(\mathbf{U}) = \gamma(\mathbf{V}) = \sqrt{\rho r/L}$, and $\xi(\mathbf{U}, \mathbf{V}) = \sqrt{\rho r}/L$ for some constant $\rho > 0$. Matrix $\mathbf{A}_0$ is also drawn from the uniform sparsity model outlined in Section VI-A. From Corollary 1 and the results in Lemmata 5 and 6, it follows that $s \precsim \frac{L^2}{\log(L)}\min\{\frac{1}{r}, \frac{L}{r^3}\}$ suffices for exact recovery with high probability. In particular, if $r \leq \sqrt{L}$, the pair $(r, s)$ should only satisfy $sr \precsim \frac{L^2}{\log(L)}$. In contrast, results in [17] only offer recovery guarantees for rank and sparsity levels up to $s\sqrt{r} \precsim \frac{L^{3/2}}{\log(L)}$, which are weaker than those derived from Corollary 1 as $r \leq L$. The results in [17] have been improved in [18, Th. 3], which allows rank and sparsity levels up to $sr \precsim \frac{L^2}{\log(L)}$ as obtained from Corollary 1. Note that Corollary 1, [17], and [18] offer *deterministic* reconstruction guarantees, where [18] yields the best results. Still in the aforementioned random setting, the condition induced from Corollary 1 is comparable with [18] thanks to the existing tight probabilistic bounds for $\mu(\Phi, \Omega)$. The results in [11] however, build on the uniform sparsity model for $\mathbf{A}_0$, and provide superior probabilistic guarantees up to $s \precsim L^2$ and $r \precsim L$.

It is worth noting that in the presence of the compression matrix $\mathbf{R}$, more stringent conditions are imposed on the rank and sparsity level, as stated in Corollary 3. This is mainly because of the dominant summand $\theta_{1,1}(\mathbf{R})[\sqrt{2}k + s\gamma^2(V)]$ in $\omega$ (cf., Theorem 1), which limits the extent to which $r$ and $s$ can be increased. If the correlation between any two columns of $\mathbf{R}$ is small, then higher rank and less sparse matrices can be exactly recovered.

## VII. Algorithms

This section deals with iterative algorithms to solve the nonsmooth convex optimization problem (P1).

### *A. APG Algorithm*

The class of APG algorithms were originally studied in [41] and [42], and they have been popularized for $\ell_1$-norm regularized regression; mostly due to the success of the fast iterative shrinkage-thresholding algorithm [4]. Recently, APG algorithms have been applied to matrix-valued problems such as those arising with nuclear-norm regularized estimators for matrix completion [49], and for (stable) PCP [34], [55]. APG algorithms offer several attractive features, most notably a convergence rate guarantee of $\mathcal{O}(1/\sqrt{\epsilon})$ iterations to return an $\epsilon$–optimal solution. In addition, APG algorithms are first-order methods that scale nicely to high-dimensional problems arising with large networks.

The algorithm developed here builds on the APG iterations in [34] proposed to solve the stable PCP problem. One can relax the equality constraint in (P1) and instead solve

$$\text{(P2)} \quad \min_{\mathbf{S}} \quad \left\{\nu\|\mathbf{X}\|_* + \nu\lambda\|\mathbf{A}\|_1 + \frac{1}{2}\|\mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{A}\|_F^2\right\}$$

with $\mathbf{S} := [\mathbf{X}', \mathbf{A}']'$, where the least-square term penalizes violations of the equality constraint, and $\nu > 0$ is a penalty coefficient. When $\nu$ approaches zero, (P2) achieves the optimal solution of (P1) [5]. The gradient of $f(\mathbf{S}) := \frac{1}{2}\|\mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{A}\|_F^2$ is Lipschitz continuous with a (minimum) Lipschitz constant

$L_f = \lambda_{\max}([\mathbf{I}_L \ \ \mathbf{R}]'[\mathbf{I}_L \ \ \mathbf{R}])$, i.e., $\|\nabla f(\mathbf{S}_1) - \nabla f(\mathbf{S}_2)\| \leq L_f\|\mathbf{S}_1 - \mathbf{S}_2\|, \forall\ \mathbf{S}_1, \mathbf{S}_2$ in the domain of $f$.

Instead of directly optimizing the cost in (P2), APG algorithms minimize a sequence of overestimators, obtained at judiciously chosen points $\mathbf{T}$. Define $g(\mathbf{S}) := \nu\|\mathbf{X}\|_* + \nu\lambda\|\mathbf{A}\|_1$ and form the quadratic approximation

$$\begin{aligned} Q(\mathbf{S},\mathbf{T}) &:= f(\mathbf{T}) + \langle \nabla f(\mathbf{T}), \mathbf{S}-\mathbf{T}\rangle + \frac{L_f}{2}\|\mathbf{S}-\mathbf{T}\|_F^2 + g(\mathbf{S}) \\ &= \frac{L_f}{2}\|\mathbf{S}-\mathbf{G}\|_F^2 + g(\mathbf{S}) + f(\mathbf{T}) - \frac{1}{2L_f}\|\nabla f(\mathbf{T})\|_F^2 \end{aligned} \tag{29}$$

where $\mathbf{G} := \mathbf{T} - (1/L_f)\nabla f(\mathbf{T})$. With $k = 1, 2, \ldots$ denoting iterations, APG algorithms generate the sequence of iterates

$$\begin{aligned} \mathbf{S}[k] &:= \arg\min_{\mathbf{S}} Q(\mathbf{S}, \mathbf{T}[k-1]) \\ &= \arg\min_{\mathbf{S}} \left\{ \frac{L_f}{2}\|\mathbf{S} - \mathbf{G}[k]\|_F^2 + g(\mathbf{S}) \right\} \end{aligned} \tag{30}$$

where the second equality follows from the fact that the last two summands in (29) do not depend on $\mathbf{S}$. There are two key aspects to the success of APG algorithms. First, is the selection of the points $\mathbf{T}[k]$ where the sequence of approximations $Q(\mathbf{S}, \mathbf{T}[k])$ are formed, since these strongly determine the algorithm's convergence rate. The choice $\mathbf{T}[k] = \mathbf{S}[k] + \frac{t[k-1]-1}{t[k]}(\mathbf{S}[k] - \mathbf{S}[k-1])$, where $t[k] = \left[1 + \sqrt{4t^2[k-1]+1}\right]/2$, has been shown to significantly accelerate the algorithm resulting in convergence rate no worse than $\mathcal{O}(1/k^2)$ [4]. The second key element stems from the possibility of efficiently solving the sequence of subproblems (30). For the particular case of (P2), note that (30) decomposes into

$$\mathbf{X}[k+1] := \arg\min_{\mathbf{X}} \left\{ \frac{L_f}{2}\|\mathbf{X} - \mathbf{G}_X[k]\|_F^2 + \nu\|\mathbf{X}\|_* \right\} \tag{31}$$

$$\mathbf{A}[k+1] := \arg\min_{\mathbf{A}} \left\{ \frac{L_f}{2}\|\mathbf{A} - \mathbf{G}_A[k]\|_F^2 + \nu\lambda\|\mathbf{A}\|_1 \right\} \tag{32}$$

where $\mathbf{G}[k] = [\mathbf{G}'_X[k] \ \ \mathbf{G}'_A[k]]'$. Letting $\mathcal{S}_\tau(\mathbf{M})$ with $(i,j)$th entry given by $\text{sign}(m_{i,j})\max\{|m_{i,j}| - \tau, 0\}$ denote the soft-thresholding operator, and $\mathbf{U\Sigma V}' = \text{svd}(\mathbf{G}_X[k])$ the singular value decomposition of matrix $\mathbf{G}_X[k]$, it follows that (see, e.g., [34])

$$\mathbf{X}[k+1] = \mathbf{U}\mathcal{S}_{\frac{\nu}{L_f}}[\mathbf{\Sigma}]\mathbf{V}', \quad \mathbf{A}[k+1] = \mathcal{S}_{\frac{\lambda\nu}{L_f}}[\mathbf{G}_A[k]]. \tag{33}$$

A continuation technique is employed to speed-up convergence of the APG algorithm. The penalty parameter $\nu$ is initialized with a large value $\nu_0$, and is decreased geometrically until it reaches the target value of $\bar{\nu}$. The APG algorithm is tabulated as Algorithm 1. Similar to [34] and [49], the iterations terminate whenever the norm of $\mathbf{Z}[k+1]$ in (34) at the bottom of the page drops below some prescribed tolerance, i.e., $\|\mathbf{Z}[k+1]\|_F \leq \text{tol} \times \max(1, L_f\|\mathbf{X}[k]\|_F)$. As detailed in [49], the quantity $\|\mathbf{Z}[k+1]\|_F$ upper bounds the distance between the origin and the set of subgradients of the cost in (P2), evaluated at $\mathbf{S}[k+1]$.

Before concluding this section, it is worth noting that Algorithm 1 has good convergence performance, and quantifiable iteration complexity as asserted in the following proposition adapted from [4] and [34].

*Proposition 2 [34]:* Let $h(.)$ and $\{\bar{\mathbf{A}}, \bar{\mathbf{X}}\}$ denote, respectively, the cost and an optimal solution of (P2) when $\nu := \bar{\nu}$. For $k > k_0 := \frac{\log(\nu_0/\bar{\nu})}{\log(1/\upsilon)}$, the iterates $\{\mathbf{A}[k], \mathbf{X}[k]\}$ generated by Algorithm 1 satisfy

$$\begin{aligned} |h(\mathbf{A}[k], \mathbf{X}[k]) - h(\bar{\mathbf{A}}, \bar{\mathbf{X}})| \leq \frac{4\|\mathbf{A}[k_0] - \bar{\mathbf{A}}\|_F^2}{(k-k_0+1)^2} \\ + \frac{4\|\mathbf{X}[k_0] - \bar{\mathbf{X}}\|_F^2}{(k-k_0+1)^2}. \end{aligned}$$

**Algorithm 1**: APG solver for (P1)

```
input $\mathbf{Y}, \mathbf{R}, \lambda, \upsilon, \nu_0, \bar{\nu}$, and $L_f = \lambda_{\max}([\mathbf{I}_L \ \mathbf{R}]'[\mathbf{I}_L \ \mathbf{R}])$
initialize $\mathbf{X}[0] = \mathbf{X}[-1] = \mathbf{0}_{L\times T}$, $\mathbf{A}[0] = \mathbf{A}[-1] = \mathbf{0}_{F\times T}$, $t[0] = t[-1] = 1$, and set $k = 0$.
while not converged do
    $\mathbf{T}_X[k] = \mathbf{X}[k] + \frac{t[k-1]-1}{t[k]}(\mathbf{X}[k] - \mathbf{X}[k-1])$.
    $\mathbf{T}_A[k] = \mathbf{A}[k] + \frac{t[k-1]-1}{t[k]}(\mathbf{A}[k] - \mathbf{A}[k-1])$.
    $\mathbf{G}_X[k] = \mathbf{T}_X[k] + \frac{1}{L_f}(\mathbf{Y} - \mathbf{T}_X[k] - \mathbf{R}\mathbf{T}_A[k])$.
    $\mathbf{G}_A[k] = \mathbf{T}_A[k] + \frac{1}{L_f}\mathbf{R}'(\mathbf{Y} - \mathbf{T}_X[k] - \mathbf{R}\mathbf{T}_A[k])$.
    $\mathbf{U\Sigma V}' = \text{svd}(\mathbf{G}_X[k])$,    $\mathbf{X}[k+1] = \mathbf{U}\mathcal{S}_{\nu[k]/L_f}(\mathbf{\Sigma})\mathbf{V}'$.
    $\mathbf{A}[k+1] = \mathcal{S}_{\lambda\nu[k]/L_f}(\mathbf{G}_A[k])$.
    $t[k+1] = \left[1 + \sqrt{4t^2[k]+1}\right]/2$
    $\nu[k+1] = \max\{\upsilon\nu[k], \bar{\nu}\}$
    $k \leftarrow k+1$
end while
return $\mathbf{X}[k]$, $\mathbf{A}[k]$
```

### B. AD-MoM Algorithm

The AD-MoM is an iterative augmented Lagrangian method especially well-suited for parallel processing [6], which has

$$\mathbf{Z}[k+1] := \begin{bmatrix} L_f(\mathbf{T}_X[k] - \mathbf{X}[k+1]) + (\mathbf{X}[k+1] + \mathbf{R}\mathbf{A}[k+1] - \mathbf{T}_X[k] - \mathbf{R}\mathbf{T}_A[k]) \\ L_f(\mathbf{T}_A[k] - \mathbf{A}[k+1]) + \mathbf{R}'(\mathbf{X}[k+1] + \mathbf{R}\mathbf{A}[k+1] - \mathbf{T}_X[k] - \mathbf{R}\mathbf{T}_A[k]) \end{bmatrix} \tag{34}$$

been proven successful to tackle the optimization tasks encountered e.g., in statistical learning problems [9], [39]. While the AD-MoM could be directly applied to (P1), $\mathbf{R}$ couples the entries of $\mathbf{A}$ and it turns out this yields more difficult $\ell_1$-norm minimization subproblems per iteration. To overcome this challenge, a common technique is to introduce an auxiliary (decoupling) variable $\mathbf{B}$, and formulate the following optimization problem

$$\text{(P3)} \quad \min_{\{\mathbf{X},\mathbf{A},\mathbf{B}\}} \|\mathbf{X}\|_* + \lambda\|\mathbf{A}\|_1$$

$$\text{s. t.} \;\; \mathbf{Y} = \mathbf{X} + \mathbf{R}\mathbf{B} \tag{35}$$

$$\mathbf{B} = \mathbf{A} \tag{36}$$

which is equivalent to (P1). To tackle (P3), associate Lagrange multipliers $\tilde{\mathbf{M}}$ and $\bar{\mathbf{M}}$ with the constraints (35) and (36), respectively. Next, introduce the quadratically *augmented* Lagrangian function

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mathbf{A}, \mathbf{B}, \tilde{\mathbf{M}}, \bar{\mathbf{M}}) = \|\mathbf{X}\|_* &+ \lambda\|\mathbf{A}\|_1 \\ &+ \langle\tilde{\mathbf{M}}, \mathbf{B} - \mathbf{A}\rangle + \langle\bar{\mathbf{M}}, \mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{B}\rangle \\ &+ \frac{c}{2}\|\mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{B}\|_F^2 + \frac{c}{2}\|\mathbf{A} - \mathbf{B}\|_F^2\end{aligned} \tag{37}$$

where $c$ is a positive penalty coefficient. Splitting the primal variables into two groups $\{\mathbf{X}, \mathbf{A}\}$ and $\{\mathbf{B}\}$, the AD-MoM solver entails an iterative procedure comprising three steps per iteration $k = 1, 2, \ldots$

**[S1] Update dual variables:**

$$\tilde{\mathbf{M}}[k] = \tilde{\mathbf{M}}[k-1] + c(\mathbf{B}[k] - \mathbf{A}[k]) \tag{38}$$

$$\bar{\mathbf{M}}[k] = \bar{\mathbf{M}}[k-1] + c(\mathbf{Y} - \mathbf{X}[k] - \mathbf{R}\mathbf{B}[k]). \tag{39}$$

**[S2] Update first group of primal variables:**

$$\begin{aligned}\mathbf{X}[k+1] = \arg\min_{\mathbf{X}} \Big\{&\frac{c}{2}\|\mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{B}[k]\|_F^2 \\ &- \langle\bar{\mathbf{M}}[k], \mathbf{X}\rangle + \|\mathbf{X}\|_*\Big\}.\end{aligned} \tag{40}$$

$$\begin{aligned}\mathbf{A}[k+1] = \arg\min_{\mathbf{A}} \Big\{&\frac{c}{2}\|\mathbf{A} - \mathbf{B}[k]\|_F^2 - \langle\tilde{\mathbf{M}}[k], \mathbf{A}\rangle \\ &+ \lambda\|\mathbf{A}\|_1\Big\}.\end{aligned} \tag{41}$$

**[S3] Update second group of primal variables:**

$$\begin{aligned}\mathbf{B}[k+1] = \arg\min_{\mathbf{B}} \Big\{&\frac{c}{2}\|\mathbf{Y} - \mathbf{X}[k+1] - \mathbf{R}\mathbf{B}\|_F^2 \\ &+ \frac{c}{2}\|\mathbf{A}[k+1] - \mathbf{B}\|_F^2 \\ &- \langle\mathbf{R}'\bar{\mathbf{M}}[k] - \tilde{\mathbf{M}}[k], \mathbf{B}\rangle\Big\}.\end{aligned} \tag{42}$$

This three-step procedure implements a block-coordinate descent on the augmented Lagrangian, with dual variable updates. The minimization (40) can be recast as (31), hence $\mathbf{X}[k+1]$ is iteratively updated through singular value thresholding. Likewise, (41) can be put in the form (32) and the entries of $\mathbf{A}[k+1]$ are updated via parallel soft-thresholding operations. Finally, (42) is a strictly convex unconstrained quadratic program, whose closed-form solution is obtained as the root of the linear equation corresponding to the first-order condition for optimality. The AD-MoM solver is tabulated under Algorithm 2. Suitable termination criteria are suggested in [9, p. 18].

Conceivably, $F$ can be quite large, thus inverting the $F \times F$ matrix $\mathbf{R}'\mathbf{R} + \mathbf{I}_F$ to update $\mathbf{B}[k+1]$ could be complex computationally. Fortunately, the inversion needs to be carried out once, and can be performed and cached offline. In addition, to reduce the inversion cost, the SVD of the compression matrix $\mathbf{R} = \mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}'_R$ can be obtained first, and the matrix inversion lemma can be subsequently employed to obtain $[\mathbf{R}'\mathbf{R} + \mathbf{I}_F]^{-1} = [\mathbf{I}_L - \mathbf{V}_R\mathbf{C}\mathbf{V}'_R]$, where $\mathbf{C} := \text{diag}\left(\frac{\sigma_1^2}{1+\sigma_1^2}, \ldots, \frac{\sigma_L^2}{1+\sigma_p^2}\right)$ and $p = \text{rank}(\mathbf{R}) \ll F$. Finally, note that the AD-MoM algorithm converges to the global optimum of the convex program (P1) as stated in the next proposition.

*Proposition 3 [6]:* For any value of the penalty coefficient $c > 0$, the iterates $\{\mathbf{X}[k], \mathbf{A}[k]\}$ converge to the optimal solution of (P1) as $k \to \infty$.

**Algorithm 2**: AD-MoM solver for (P1)

```
input $\mathbf{Y}, \mathbf{R}, \lambda, c$
initialize $\mathbf{X}[0] = \bar{\mathbf{M}}[-1] = \mathbf{0}_{L\times T}$, $\mathbf{A}[0] = \mathbf{B}[0] = \tilde{\mathbf{M}}[-1] = \mathbf{0}_{F\times T}$, and set $k = 0$.
while not converged do
    [S1] Update dual variables:
    $\tilde{\mathbf{M}}[k] = \tilde{\mathbf{M}}[k-1] + c(\mathbf{B}[k] - \mathbf{A}[k])$
    $\bar{\mathbf{M}}[k] = \bar{\mathbf{M}}[k-1] + c(\mathbf{Y} - \mathbf{X}[k] - \mathbf{R}\mathbf{B}[k])$
    [S2] Update first group of primal variables:
    $\mathbf{U}\mathbf{\Sigma}\mathbf{V}' = \text{svd}(\mathbf{Y} - \mathbf{R}\mathbf{B}[k] + c^{-1}\bar{\mathbf{M}}[k])$,
    $\mathbf{X}[k+1] = \mathbf{U}\mathcal{S}_{1/c}(\mathbf{\Sigma})\mathbf{V}'$.
    $\mathbf{A}[k+1] = c^{-1}\mathcal{S}_{\lambda}(\tilde{\mathbf{M}}[k] + c\mathbf{B}[k])$.
    [S3] Update second group of primal variables:
    $\mathbf{B}[k+1] = \mathbf{A}[k+1] + (\mathbf{R}'\mathbf{R} + \mathbf{I}_F)^{-1}\Big[\mathbf{R}'(\mathbf{Y} - \mathbf{X}[k+1] - \mathbf{R}\mathbf{A}[k+1]) - c^{-1}(\tilde{\mathbf{M}}[k] - \mathbf{R}'\bar{\mathbf{M}}[k])\Big]$
    $k \leftarrow k+1$
end while
return $\mathbf{A}[k], \mathbf{X}[k]$
```

Before moving on to performance evaluation, a couple of remarks are in order.

*Remark 7 (Tradeoff Between Stability and Convergence Rate):* The APG algorithm exhibits a convergence rate guarantee of $\mathcal{O}(1/k^2)$ [41], while AD-MoM only attains $\mathcal{O}(1/k)$ [27]. For the problem considered here, APG needs an appropriate continuation technique to achieve the predicted performance [34]. Extensive numerical tests with Algorithm 1 suggest that the convergence rate can vary considerably for different choices, e.g., of the matrix $\mathbf{R}$. The AD-MoM algorithm on the other hand exhibits less variability in terms of performance and only requires tuning $c$. It is also better suited for the constrained formulation (P1), since it does not need to resort to a relaxation.

*Remark 8 (Distributed Algorithms):* In the anomaly detection context outlined in Section II-A, implementing Algorithms 1 and 2 presume that network nodes communicate their local link traffic measurements to a central monitoring station, which uses their aggregation in $\mathbf{Y}$ to unveil anomalies. While for the most part this is the prevailing operational paradigm adopted in current networks, there are limitations associated with this architecture. For instance, fusing all this information may entail excessive communication overhead. Moreover, minimizing the exchanges of raw measurements may be desirable to reduce unavoidable communication errors that translate to noise and missing data. Performing the optimization in a centralized fashion also raises robustness concerns, since the central monitoring station represents an isolated point of failure. These reasons motivate devising *fully distributed* algorithms for unveiling anomalies in large scale networks, whereby each node carries out simple computational tasks locally, relying only on its local measurements and messages exchanged with its directly connected neighbors. This is the subject dealt with in an algorithmic companion paper [37], which puts forth a general framework for in-network sparsity-regularized rank minimization.

## VIII. Performance Evaluation

The performance of (P1) is assessed in this section via computer simulations.

*Selection of Tuning Parameters:* Theorem 1 provides a range of parameters $\lambda \in (\lambda_{\min}, \lambda_{\max})$ such that (P1) exactly recovers $\{\mathbf{X}_0, \mathbf{A}_0\}$ in (1). However, it may be infeasible to compute $\{\lambda_{\min}, \lambda_{\max}\}$, since they depend on e.g., $\delta_k(\mathbf{R})$ which is NP-hard to evaluate [15]. Besides, in practice, the observations (1) are typically contaminated with noise $\mathbf{E} \in \mathbb{R}^{L\times T}$ [cf., (3)]. To account for the noise, the optimization problem (P2) is considered, where for convenience the sparsity and rank-controlling parameters are redefined here as $\lambda_1 := \nu\lambda$ and $\lambda_* := \nu$, respectively. To tune $\{\lambda_1, \lambda_*\}$, a simple strategy is to optimize the relative error $\|\hat{\mathbf{A}} - \mathbf{A}_0\|_F/\|\mathbf{A}_0\|_F$, with $\mathbf{A}_0$ and $\hat{\mathbf{A}}$ denoting the true and estimated sparse matrices, respectively. In particular, one needs to perform a grid search over the bounded two-dimensional region $\mathcal{R} := \{(\lambda_1, \lambda_*) : \lambda_1 \in (0, \|\mathbf{R}'\mathbf{Y}\|_\infty],\ \lambda_* \in (0, \|\mathbf{Y}\|]\}$. The corresponding bounds are derived from the optimality conditions for (P2), which indicate that for $(\lambda_1, \lambda_*) \in \mathcal{R}^c$ the optimal solution is $\{\mathbf{0}_{L\times T}, \mathbf{0}_{F\times T}\}$.

Practical rules that do not require knowledge of $\mathbf{A}_0$ can be devised along the lines of [2] and [12]. Supposing that the true values are zero, choosing $\lambda_1 > \|\mathbf{R}'\mathbf{E}\|_\infty$ and $\lambda_* > \|\mathbf{E}\|$, the estimator (P2) outputs $\{\hat{\mathbf{X}} = \mathbf{0}_{L\times T}, \hat{\mathbf{A}} = \mathbf{0}_{F\times T}\}$. In general, this choice mitigates noise, but it may overshrink the true values. To avoid overshrinking, these parameters can be chosen close to their corresponding lower bounds, e.g., pick $\lambda_* = \|\mathbf{E}\|$ and $\lambda_1 = \|\mathbf{R}'\mathbf{E}\|_\infty$. One can further simplify the candidate parameters by making the following reasonable assumptions: i) Gaussian noise $e_{l,t} \sim \mathcal{N}(0, \sigma^2)$, and ii) large dimensions $F, T \to \infty$. It is then known that $(\sqrt{F} + \sqrt{T})^{-1}\|\mathbf{E}\| \to \sigma$, almost surely, see, e.g., [12], and thus, one can pick $\lambda_* = (\sqrt{F} + \sqrt{T})\sigma$. Also, large deviation tail bounding implies that $\|\mathbf{R}'\mathbf{E}\|_\infty \leq 4\sigma \max_i \|\mathbf{R}\mathbf{e}_i\|_2 \log(FT)$ with high probability, which suggests selecting $\lambda_1 = \sigma \max_i \|\mathbf{R}\mathbf{e}_i\|_2 \log(FT)$. Notice that in the noiseless case ($\sigma = 0$) one can pick $\lambda = \lambda_1/\lambda_* = \max_i \|\mathbf{R}\mathbf{e}_i\| \log(FT)/(\sqrt{F} + \sqrt{T})$.

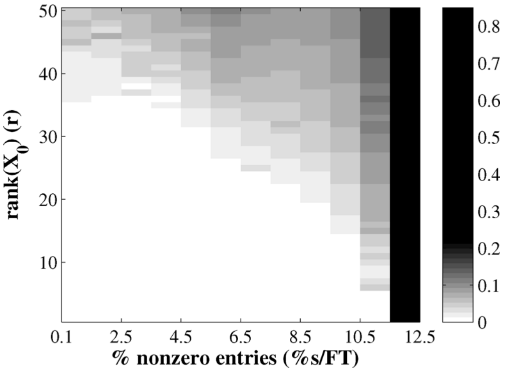

Fig. 1. Relative error $e_r := \|\mathbf{A}_0 - \hat{\mathbf{A}}\|_F/\|\mathbf{A}_0\|_F$ for various values of $r$ and $s$, where $L = 105$, $F = 210$, and $T = 420$. White represents exact recovery ($e_r \approx 0$), while black represents $e_r \approx 1$.

### A. Exact Recovery

Data matrices are generated according to $\mathbf{Y} = \mathbf{X}_0 + \mathbf{V}_R'\mathbf{A}_0$. The low-rank component $\mathbf{X}_0$ is generated from the bilinear factorization model $\mathbf{X}_0 = \mathbf{V}_R'\mathbf{W}\mathbf{Z}'$, where $\mathbf{W}$ and $\mathbf{Z}$ are $L\times r$ and $T\times r$ matrices with i.i.d. entries drawn from Gaussian distributions $\mathcal{N}(0, 1/L)$ and $\mathcal{N}(0, 1/T)$, respectively. Every entry of $\mathbf{A}_0$ is randomly drawn from the set $\{-1, 0, 1\}$ with $\Pr(a_{i,j} = -1) = \Pr(a_{i,j} = 1) = \pi/2$. The columns of $\mathbf{V}_R \in \mathbb{R}^{F\times L}$ comprise the right singular vectors of the random matrix $\mathbf{R} = \mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}_R'$, with i.i.d. Bernoulli entries with parameter $1/2$ (cf., Remark 2). The dimensions are $L = 105$, $F = 210$, and $T = 420$. To demonstrate that (P1) is capable of recovering the exact values of $\{\mathbf{X}_0, \mathbf{A}_0\}$, the optimization problem is solved for a wide range of values of $r$ and $s$ using the APG algorithm (cf., Algorithm 1).

Let $\hat{\mathbf{A}}$ denote the solution of (P1) for a suitable value of $\lambda$. Fig. 1 depicts the relative error in recovering $\mathbf{A}_0$, namely $\|\hat{\mathbf{A}} - \mathbf{A}_0\|_F/\|\mathbf{A}_0\|_F$ for various values of $r$ and $s$. It is apparent that (P1) succeeds in recovering $\mathbf{A}_0$ for sufficiently sparse $\mathbf{A}_0$ and low-rank $\mathbf{X}_0$ from the observed data $\mathbf{Y}$. Interestingly, in cases such as $s = 0.1 \times FT$ or $r = 0.3 \times \min(L, T)$, there is hope for recovery. In this example, one can exactly recover $\{\mathbf{X}_0, \mathbf{A}_0\}$ when $s = 0.0127\times FT$ and $r = 0.2381\times \min(L, T)$. A similar trend is observed for the recovery of $\mathbf{X}_0$, and the corresponding plot is omitted to avoid unnecessary repetition. For different sizes of the matrix $\mathbf{R}$, performance results averaged over ten realizations of the experiment are listed in Table I. The smaller the compression ratio $L/F$ becomes, less observations are available and performance degrades accordingly. In particular, the error performance degrades significantly for a challenging instance where $L/F = 0.2$ and $r = 0.4 \times \min(L, F)$ (cf., the last row of Table I).

The results of [11] and [17] assert that exact recovery of $\{\mathbf{X}_0, \mathbf{A}_0\}$ from the observations $\mathbf{Y} = \mathbf{X}_0 + \mathbf{A}_0$ is possible

TABLE I
RECOVERY PERFORMANCE BY VARYING THE SIZE OF $\mathbf{R}$ WHEN $r = 10$ AND $\pi = 0.05$

| $L$ | $\text{rank}(\mathbf{X}_0)$ | $\|\mathbf{A}_0\|_0$ | $\text{rank}(\hat{\mathbf{X}})$ | $\|\hat{\mathbf{A}}\|_0$ | $\|\hat{\mathbf{A}} - \mathbf{A}_0\|_F / \|\mathbf{A}_0\|_F$ |
|---|---|---|---|---|---|
| $F$ | 10 | 4410 | 10 | 4419 | $2.0809 \times 10^{-6}$ |
| $F/2$ | 10 | 4410 | 10 | 4407 | $6.4085 \times 10^{-5}$ |
| $F/3$ | 10 | 4410 | 10 | 9365 | $7.76 \times 10^{-2}$ |
| $F/5$ | 10 | 4410 | 14 | 14690 | $6.331 \times 10^{-1}$ |

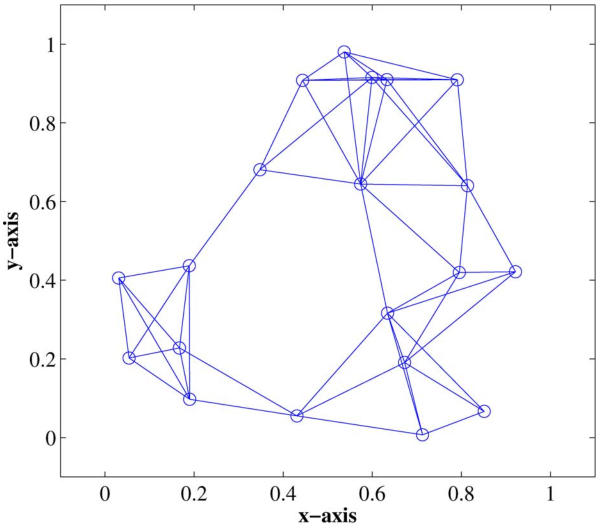

Fig. 2. Network topology graph.

under some technical conditions. Even though the algorithms therein are not directly applicable here due to the presence of $\mathbf{R}$, one may still consider applying PCP after suitable preprocessing of $\mathbf{Y}$. One possible approach is to find the LS estimate of the superposition $\mathbf{X}_0 + \mathbf{A}_0$ as $\hat{\mathbf{Y}} = \mathbf{R}^{\dagger}\mathbf{Y}$, and then feed a PCP algorithm with $\hat{\mathbf{Y}}$ to obtain $\{\mathbf{X}_0, \mathbf{A}_0\}$. Comparisons between (P1) and the aforesaid two-step procedure are summarized in Table II. It is apparent that the heuristic performs very poorly, which is mainly due to the null space of matrix $\mathbf{R}$ (when $F = 2L$) that renders LS estimation inaccurate.

### B. Unveiling Network Anomalies

*Synthetic Network Data:* A network of $N = 20$ agents is considered as a realization of the random geometric graph model, i.e., agents are randomly placed on the unit square and two agents communicate with each other if their Euclidean distance is less than a prescribed communication range of 0.35; see Fig. 2. The network graph is bidirectional and comprises $L = 106$ links, and $F = N(N-1) = 380$ OD flows. For each candidate OD pair, minimum hop count routing is considered to form the routing matrix $\mathbf{R}$. With $r = 10$, matrices $\{\mathbf{X}_0, \mathbf{A}_0\}$ are generated as explained in Section VIII-A. With reference to (2), the entries of $\mathbf{E}$ are i.i.d., zero-mean, Gaussian with variance $\sigma^2$, i.e., $e_{l,t} \sim \mathcal{N}(0, \sigma^2)$.

*Real Network Data:* Real data including OD flow traffic levels are collected from the operation of the Internet2 network (Internet backbone network across USA) [1]. OD flow traffic levels are recorded for a three-week operation of Internet2 during Dec. 8–28, 2008 [32]. Internet2 comprises $N = 11$ nodes, $L = 41$ links, and $F = 121$ flows. Given the OD flow traffic measurements, the link loads in $\mathbf{Y}$ are obtained through multiplication with the Internet2 routing matrix [1]. Even though $\mathbf{Y}$ is "constructed" here from flow measurements, link loads can be typically acquired from simple network management protocol traces [48]. The available OD flows are a superposition of "clean" and anomalous traffic, i.e., the sum of unknown "ground-truth" low-rank and sparse matrices $\mathbf{X}_0 + \mathbf{A}_0$ adhering to (2) when $\mathbf{R} = \mathbf{I}_L$. Therefore, PCP is applied first to obtain an estimate of the "ground-truth" $\{\mathbf{X}_0, \mathbf{A}_0\}$. The estimated $\mathbf{X}_0$ exhibits three dominant singular values, confirming the low-rank property of $\mathbf{X}_0$.

*Comparison With the PCA-Based Method:* To highlight the merits of the proposed anomaly detection algorithm, its performance is compared with the workhorse PCA-based approach of [32]. The crux of this method is that the anomaly-free data is expected to be low-rank, whereas the presence of anomalies considerably increases the rank of $\mathbf{Y}$. PCA requires *a priori* knowledge of the rank of the anomaly-free traffic matrix and is unable to identify multiple anomalous flows, i.e., the scope of [32] is limited to a single anomalous flow per time slot. Different from [32], the developed framework here enables identifying multiple anomalous flows per time instant. To assess performance, the detection rate will be used as figure of merit, which measures the algorithm's success in identifying anomalies across both flows and time.

For the synthetic data case, ROC curves are depicted in Fig. 3 (top), for different values of the rank required to run the PCA-based method. It is apparent that the proposed scheme detects accurately the anomalies, even at low false alarm rates. For the particular case of $P_F = 10^{-4}$ and $P_D = 0.97$, Fig. 3 (bottom) illustrates the magnitude of the true and estimated anomalies across flows and time. Similar results are depicted for the Internet2 data in Fig. 4, where it is also apparent that the proposed method markedly outperforms PCA in terms of detection performance. For an instance of $P_F = 0.04$ and $P_D = 0.93$, Fig. 4 (bottom) shows the effectiveness of the proposed algorithm in terms of unveiling the anomalous flows and time instants.

*Remark 9 (Incoherence Conditions):* For the matrices involved in the anomaly detection problem, some of the incoherence conditions required by Theorem 1 may not hold. For instance, with $\mathbf{X}_0 = \mathbf{R}\mathbf{Z}_0$ [cf., (2)], quantity $\gamma_R(\mathbf{U})$ may not be small enough. In addition, it is challenging to find binary $\{0, 1\}$ routing matrices with desirable RICs. Still, the conditions in Theorem 1 are only sufficient and the numerical tests in this section demonstrate that the proposed algorithm performs well in practice. This observation naturally motivates follow-up research aimed at closing this gap between theory and practice.

## IX. CLOSING COMMENTS

This paper deals with recovery of low-rank plus *compressed* sparse matrices via convex optimization. The corresponding task arises with network traffic monitoring, dynamic MRI, and singing voice separation from music accompaniment, while it encompasses compressive sampling and principal components pursuit. To estimate the unknowns, a convex optimization program is formulated that minimizes a tradeoff between the nuclear and $\ell_1$-norm of the low-rank and sparse components, respectively, subject to a data modeling constraint. A deterministic approach is adopted to characterize local identifiability and

TABLE II
PERFORMANCE COMPARISON OF LS-PCP AND ALGORITHM 1 AVERAGED OVER TEN RANDOM REALIZATIONS

| Algorithm | $r=5,\ \pi=0.01$ | $r=5,\ \pi=0.05$ | $r=10,\ \pi=0.01$ | $r=10,\ \pi=0.05$ |
|---|---|---|---|---|
| LS-PCP | 0.6901 | 0.6975 | 0.7001 | 0.7023 |
| Algorithm 1 | $7.81\times10^{-6}$ | $3.037\times10^{-5}$ | $1.69\times10^{-5}$ | $6.4\times10^{-5}$ |

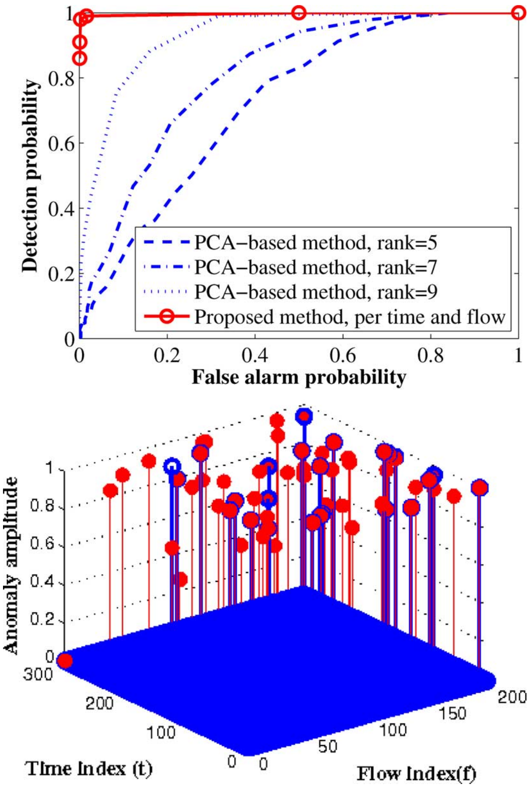

Fig. 3. Performance for synthetic data. (Top) ROC curves of the proposed versus the PCA-based method with $\pi=0.001$, $r=10$ and $\sigma=0.1$. (Bottom) Amplitude of the true and estimated anomalies for $P_F=10^{-4}$ and $P_D=0.97$. Lines with open and filled circle markers denote the true and estimated anomalies, respectively.

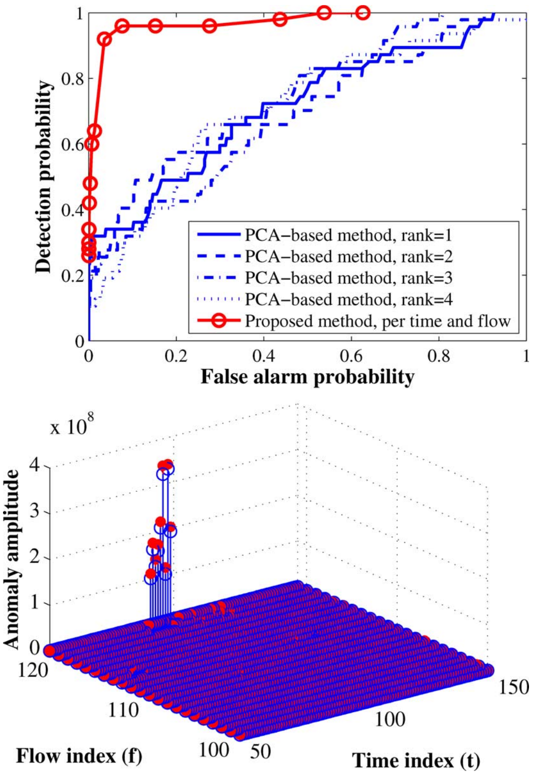

Fig. 4. Performance for Internet2 network data. (Top) ROC curves of the proposed versus the PCA-based method. (Bottom) Amplitude of the true and estimated anomalies for $P_F=0.04$ and $P_D=0.93$. Lines with open and filled circle markers denote the true and estimated anomalies, respectively.

sufficient conditions for exact recovery via the aforementioned convex program. Intuitively, the obtained conditions require: i) incoherent, sufficiently low-rank and sparse components; and ii) a compression matrix that behaves like an isometry when operating on sparse vectors. Because these conditions are in general NP-hard to check, it is shown that matrices drawn from certain random ensembles can be recovered with high probability. First-order iterative algorithms are developed to solve the nonsmooth optimization problem, which converge to the globally optimal solution with quantifiable complexity. Numerical tests with synthetic and real network data corroborate the effectiveness of the novel approach in unveiling traffic anomalies across flows and time.

One can envision several extensions to this work, which provide new and challenging directions for future research. For instance, it seems that the requirement of an orthonormal compression matrix is only a restriction imposed by the method of proof utilized here. There should be room for tightening the bounds used in the process of constructing the dual certificate, and hence obtain milder conditions for exact recovery. Building on [18] and [55], it would also be interesting to study stability of the proposed estimator in the presence of noise and missing data. In addition, one is naturally tempted to search for a broader class of matrices satisfying the exact recovery conditions, including e.g., non block-diagonal and binary routing (compression) matrices arising with the network anomaly detection task.

## APPENDIX

### A. Proof of Lemma 2

Suppose $\{\mathbf{X}_0,\mathbf{A}_0\}$ is an optimal solution of (P1). For the nuclear norm and the $\ell_1$-norm at point $\{\mathbf{X}_0,\mathbf{A}_0\}$, pick the subgradients $\mathbf{U}\mathbf{V}'+\mathbf{W}_0$ and $\text{sign}(\mathbf{A}_0)+\mathbf{F}_0$, respectively, satisfying the optimality condition

$$\lambda\text{sign}(\mathbf{A}_0)+\lambda\mathbf{F}=\mathbf{R}'(\mathbf{U}\mathbf{V}'+\mathbf{W}). \tag{43}$$

Consider a feasible solution $\{\mathbf{X}_0+\mathbf{R}\mathbf{H},\mathbf{A}_0-\mathbf{H}\}$ for arbitrary nonzero $\mathbf{H}$. The subgradient inequality yields

$$\begin{aligned}\|\mathbf{X}_0+\mathbf{R}\mathbf{H}\|_*+\lambda\|\mathbf{A}_0-\mathbf{H}\|\geq\|\mathbf{X}_0\|_*+\lambda\|\mathbf{A}_0\|_1&\\ +\underbrace{\langle\mathbf{U}\mathbf{V}'+\mathbf{W}_0,\mathbf{R}\mathbf{H}\rangle-\lambda\langle\text{sgn}(\mathbf{A}_0)+\mathbf{F}_0,\mathbf{H}\rangle}_{:=\varphi(\mathbf{H})}&.\end{aligned}$$

To guarantee uniqueness, $\varphi(\mathbf{H})$ must be positive. Rearranging terms, one obtains

$$\varphi(\mathbf{H}) = \langle \mathbf{W}_0, \mathbf{RH}\rangle - \lambda\langle \mathbf{F}_0, \mathbf{H}\rangle + \langle \mathbf{R}'\mathbf{U}\mathbf{V}' - \lambda \mathrm{sign}(\mathbf{A}_0), \mathbf{H}\rangle. \tag{44}$$

The value of $\mathbf{W}_0$ can be chosen such that $\langle \mathbf{W}_0, \mathbf{RH}\rangle = \|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_*$. This is because, $\|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_* = \sup_{\|\bar{\mathbf{W}}\|\leq 1} |\langle \bar{\mathbf{W}}, \mathcal{P}_{\Phi^\perp}(\mathbf{RH})\rangle|$; thus, there exists a $\bar{\mathbf{W}}$ such that $\langle \mathcal{P}_{\Phi^\perp}(\bar{\mathbf{W}}), \mathbf{RH}\rangle = \|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_*$. One can then choose $\mathbf{W}_0 := \mathcal{P}_{\Phi^\perp}(\bar{\mathbf{W}})$ since $\|\mathcal{P}_{\Phi^\perp}(\bar{\mathbf{W}})\| \leq \|\bar{\mathbf{W}}\| \leq 1$ and $\mathcal{P}_{\Phi}(\mathbf{W}_0) = \mathbf{0}_{L\times T}$. Similarly, if one selects $\mathbf{F}_0 := -\mathcal{P}_{\Omega^\perp}(\mathrm{sign}(\mathbf{H}))$, which satisfies $\mathcal{P}_{\Omega}(\mathbf{F}_0) = \mathbf{0}_{F\times T}$ and $\|\mathbf{F}_0\|_\infty = 1$, then $\langle \mathbf{F}_0, \mathbf{H}\rangle = -\|\mathcal{P}_{\Omega^\perp}(\mathbf{H})\|_1$. Now, using (43), (44) is expressed as

$$\varphi(\mathbf{H}) = \|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\| + \lambda\|\mathcal{P}_{\Omega^\perp}(\mathbf{H})\| + \langle \lambda\mathbf{F} - \mathbf{R}'\mathbf{W}, \mathbf{H}\rangle.$$

From the triangle inequality $|\langle \lambda\mathbf{F} - \mathbf{R}'\mathbf{W}, \mathbf{H}\rangle| \leq \lambda|\langle \mathbf{F}, \mathbf{H}\rangle| + |\langle \mathbf{R}'\mathbf{W}, \mathbf{H}\rangle|$, it thus follows that

$$\begin{aligned}\varphi(\mathbf{H}) \geq\ & (\|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_* - |\langle \mathbf{R}'\mathbf{W}, \mathbf{H}\rangle|) \\ & + \lambda\left(\|\mathcal{P}_{\Omega^\perp}(\mathbf{H})\|_1 - |\langle \mathbf{F}, \mathbf{H}\rangle|\right).\end{aligned} \tag{45}$$

Since $\mathcal{P}_{\Phi^\perp}(\mathbf{W}) = \mathbf{W}$, it is deduced that $|\langle \mathbf{W}, \mathbf{RH}\rangle| = |\langle \mathbf{W}, \mathcal{P}_{\Phi^\perp}(\mathbf{RH})\rangle| \leq \|\mathbf{W}\|\|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_*$. Likewise, $\mathcal{P}_{\Omega^\perp}(\mathbf{F}) = \mathbf{F}$ yields $|\langle \mathbf{F}, \mathbf{H}\rangle| = |\langle \mathbf{F}, \mathcal{P}_{\Omega^\perp}(\mathbf{H})\rangle| \leq \|\mathbf{F}\|_\infty\|\mathcal{P}_{\Omega^\perp}(\mathbf{H})\|_1$. As a result

$$\begin{aligned}\varphi(\mathbf{H}) &\geq (1-\|\mathbf{W}\|)\|\mathcal{P}_{\Phi}(\mathbf{RH})\|_* + \lambda(1-\|\mathbf{F}\|_\infty)\|\mathcal{P}_{\Omega^\perp}(\mathbf{H})\|_1 \\ &\geq (1 - \max\{\|\mathbf{W}\|, \|\mathbf{F}\|_\infty\}) \\ &\quad \times \{\|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}(\mathbf{H})\|_1\}.\end{aligned} \tag{46}$$

Now, if $\|\mathbf{W}\| < 1$ and $\|\mathbf{F}\|_\infty < 1$, since $\Phi \cap \Omega_R = \{\mathbf{0}_{L\times T}\}$ and $\mathbf{RH} \neq \mathbf{0}_{L\times T}$, $\forall \mathbf{H} \in \Omega\backslash\{\mathbf{0}_{F\times T}\}$, there is no $\mathbf{H} \in \Omega$ for which $\mathbf{RH} \in \Phi$, and therefore, $\varphi(\mathbf{H}) > 0$.

Since $\mathbf{W}$ and $\mathbf{F}$ are related through (43), upon defining $\mathbf{\Gamma} := \mathbf{R}'(\mathbf{U}\mathbf{V}' + \mathbf{W})$, which is indeed the dual variable for (P1), one can arrive at conditions C1)–C4). ■

### *B. Proof of Lemma 3*

To establish that the rows of $\mathbf{A}_\Omega$ are linearly independent, it suffices to show that $\|\mathbf{A}'\mathrm{vec}(\mathbf{H})\| > 0$, for all nonzero $\mathbf{H} \in \Omega$. It is then possible to bound

$$\begin{aligned}\|\mathbf{A}'\mathrm{vec}(\mathbf{H})\| &= \|(\mathbf{I} - \mathbf{P}_V) \otimes (\mathbf{I} - \mathbf{P}_U)\mathbf{R}\mathrm{vec}(\mathbf{H})\| \\ &= \|(\mathbf{I} - \mathbf{P}_U)\mathbf{RH}(\mathbf{I} - \mathbf{P}_V)\|_F \\ &= \|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_F = \|\mathbf{RH} - \mathcal{P}_{\Phi}(\mathbf{RH})\|_F \\ &\overset{(a)}{\geq} \|\mathbf{RH}\|_F - \|\mathcal{P}_{\Phi}(\mathbf{RH})\|_F \\ &\overset{(b)}{\geq} \|\mathbf{RH}\|_F(1 - \mu(\Omega_R, \Phi))\end{aligned} \tag{47}$$

where (a) follows from the triangle inequality, and (b) from (6). The assumption $\delta_k(\mathbf{R}) < 1$ along with the fact that no column of $\mathbf{H}$ has more than $k$ nonzero elements, imply that $\mathbf{RH} \neq \mathbf{0}_{L\times T}$. Since $\mu(\Omega_r, \Phi) < 1$ by assumption, the claim follows from (47).

To arrive at the desired bound on $\sigma_{\min}(\mathbf{A}'_\Omega)$, recall the definition of the minimum singular value [28]

$$\begin{aligned}\sigma_{\min}(\mathbf{A}'_\Omega) &= \min_{\mathbf{H}\in\Omega\backslash\{\mathbf{0}_{F\times T}\}} \frac{\|\mathbf{A}'\mathrm{vec}(\mathbf{H})\|}{\|\mathrm{vec}(\mathbf{H})\|} \\ &= \min_{\mathbf{H}\in\Omega\backslash\{\mathbf{0}_{F\times T}\}} \frac{\|(\mathbf{I} - \mathbf{P}_U)\mathbf{RH}(\mathbf{I} - \mathbf{P}_V)\|_F}{\|\mathbf{H}\|_F} \\ &\overset{(c)}{=} \min_{\mathbf{H}\in\Omega\backslash\{\mathbf{0}_{F\times T}\}} \frac{\|\mathbf{RH}\|_F}{\|\mathbf{H}\|_F} \times \frac{\|\mathcal{P}_{\Phi^\perp}(\mathbf{RH})\|_F}{\|\mathbf{RH}\|_F} \\ &\overset{(d)}{\geq} c^{1/2}(1 - \delta_k(\mathbf{R}))^{1/2} \min_{\mathbf{Z}\in\Omega_R\backslash\{\mathbf{0}_{L\times T}\}} \frac{\|\mathcal{P}_{\Phi^\perp}(\mathbf{Z})\|_F}{\|\mathbf{Z}\|_F} \\ &= c^{1/2}(1 - \delta_k(\mathbf{R}))^{1/2} \min_{\mathbf{Z}\in\Omega_R\backslash\{\mathbf{0}\}} \frac{\|\mathbf{Z} - \mathcal{P}_{\Phi}(\mathbf{Z})\|_F}{\|\mathbf{Z}\|_F} \\ &\overset{(e)}{\geq} c^{1/2}(1-\delta_k(\mathbf{R}))^{1/2}\left(1 - \max_{\mathbf{Z}\in\Omega_R\backslash\{\mathbf{0}\}} \frac{\|\mathcal{P}_{\Phi}(\mathbf{Z})\|_F}{\|\mathbf{Z}\|_F}\right) \\ &\overset{(f)}{=} c^{1/2}(1 - \delta_k(\mathbf{R}))^{1/2}(1 - \mu(\Phi, \Omega_R)).\end{aligned}$$

In obtaining (c), the assumption $\delta_k(\mathbf{R}) < 1$ along with the fact that no column of $\mathbf{H}$ has more than $k$ nonzero elements was used to ensure that $\mathbf{RH} \neq \mathbf{0}_{L\times T}$. In addition, (d) and (f) follow from the definitions (7) and (6), respectively, while (e) follows from the triangle inequality. ■

### *C. Proof of Lemma 4*

Toward establishing the first bound, from the submultiplicative property of the spectral norm, one obtains

$$\|\mathbf{Q}\| = \|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\left(\mathbf{A}_\Omega\mathbf{A}'_\Omega\right)^{-1}\| \leq \|\mathbf{A}_{\Omega^\perp}\|\|\mathbf{A}'_\Omega\left(\mathbf{A}_\Omega\mathbf{A}'_\Omega\right)^{-1}\|. \tag{48}$$

Next, upper bounds are derived for both factors on the right-hand side of (48). First, using the fact that $\mathbf{A}'\mathbf{A} = \mathbf{A}'_\Omega\mathbf{A}_\Omega + \mathbf{A}'_{\Omega^\perp}\mathbf{A}_{\Omega^\perp}$ one arrives at

$$\begin{aligned}\|\mathbf{A}_{\Omega^\perp}\|^2 &= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}'_{\Omega^\perp}\mathbf{A}_{\Omega^\perp}\mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'(\mathbf{A}'\mathbf{A} - \mathbf{A}'_\Omega\mathbf{A}_\Omega)\mathbf{x}}{\|\mathbf{x}\|^2} \\ &\leq \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x}}{\|\mathbf{x}\|^2} - \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}'_\Omega\mathbf{A}_\Omega\mathbf{x}}{\|\mathbf{x}\|^2} \\ &= \|\mathbf{A}\|^2 - \sigma^2_{\min}(\mathbf{A}'_\Omega).\end{aligned} \tag{49}$$

Note that $\mathbf{A}'_\Omega\left(\mathbf{A}_\Omega\mathbf{A}'_\Omega\right)^{-1}$ is the pseudo-inverse of the full row rank matrix $\mathbf{A}_\Omega$ (cf., Lemma 3), and thus, $\|\mathbf{A}'_\Omega\left(\mathbf{A}_\Omega\mathbf{A}'_\Omega\right)^{-1}\| = \sigma^{-1}_{\min}(\mathbf{A}'_\Omega)$ [28]. Substituting these two bounds into (48) yields

$$\|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\left(\mathbf{A}_\Omega\mathbf{A}'_\Omega\right)^{-1}\| \leq \left\{\left(\frac{\|\mathbf{A}\|}{\sigma_{\min}(\mathbf{A}'_\Omega)}\right)^2 - 1\right\}^{1/2}. \tag{50}$$

In addition, it holds that

$$\begin{aligned}\|\mathbf{A}\|^2 &= \lambda_{\max}\left\{(\mathbf{I} - \mathbf{P}_V) \otimes \mathbf{R}'(\mathbf{I} - \mathbf{P}_U)\mathbf{R}\right\} \\ &= \lambda_{\max}\{(\mathbf{I} - \mathbf{P}_V)\} \times \lambda_{\max}\left\{\mathbf{R}'(\mathbf{I} - \mathbf{P}_U)\mathbf{R}\right\} \\ &\overset{(a)}{=} \|\mathbf{R}'(\mathbf{I} - \mathbf{P}_U)\|^2 \overset{(b)}{=} 1,\end{aligned} \tag{51}$$

where in (a) and (b), it was used that the rows of $\mathbf{R}$ are orthonormal, and the maximum singular value of a projection matrix is one. Substituting (51) and the bound of Lemma 3 into (50) leads to (4).

In order to prove the second bound, first suppose that $\|\mathbf{I}-\mathbf{A}_\Omega\mathbf{A}'_\Omega\|_{\infty,\infty}<1$. Then, one can write

$$\begin{aligned}\|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\left(\mathbf{A}_\Omega\mathbf{A}'_\Omega\right)^{-1}\|_{\infty,\infty} &\leq \|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\|_{\infty,\infty}\\ &\quad\times\|\left(\mathbf{I}-(\mathbf{I}-\mathbf{A}_\Omega\mathbf{A}'_\Omega)\right)^{-1}\|_{\infty,\infty}\\ &\leq \frac{\|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\|_{\infty,\infty}}{1-\|\mathbf{I}-\mathbf{A}_\Omega\mathbf{A}'_\Omega\|_{\infty,\infty}}.\end{aligned}\tag{52}$$

In what follows, separate upper bounds are derived for $\|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\|_{\infty,\infty}$ and $\|\mathbf{I}-\mathbf{A}_\Omega\mathbf{A}'_\Omega\|_{\infty,\infty}$. For notational convenience, introduce $\mathcal{S}:=\mathrm{supp}(\mathbf{A}_0)$ (respectively, $\bar{\mathcal{S}}$ denotes the set complement). Starting with the numerator in the right-hand side of (52)

$$\begin{aligned}\|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\|_{\infty,\infty} &= \max_i\|\mathbf{e}'_i\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\|_1\\ &= \max_i\sum_k|\langle\mathbf{e}'_i\mathbf{A}_{\Omega^\perp},\mathbf{e}'_k\mathbf{A}_\Omega\rangle|\\ &= \max_j\sum_\ell|\langle\mathbf{e}'_j\mathbf{A},\mathbf{e}'_\ell\mathbf{A}\rangle|\\ &= \max_{(j_1,j_2)\in\bar{\mathcal{S}}}\sum_{(\ell_1,\ell_2)\in\mathcal{S}}g(j_1,j_2,\ell_1,\ell_2)\end{aligned}\tag{53}$$

where $g(j_1,j_2,\ell_1,\ell_2) := |\langle\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2}(\mathbf{I}-\mathbf{P}_V),(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2}\rangle|$. Following some manipulations, the summands in (53) can be expressed as

$$\begin{aligned}g(j_1,j_2,\ell_1,\ell_2) &= |\langle\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2},(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2}\rangle\\ &\quad-\langle\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2}\mathbf{P}_V,(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}\mathbf{e}'_{\ell_2}\rangle|\\ &= |\langle\mathbf{e}'_{j_2}\mathbf{e}_{\ell_2},\mathbf{e}'_{j_1}\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}\rangle\\ &\quad-\langle\mathbf{e}'_{j_2}\mathbf{P}_V\mathbf{e}_{\ell_2},\mathbf{e}'_{j_1}\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}\rangle|\\ &= |\mathbf{e}'_{j_1}\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}\mathbb{1}_{\{j_2=\ell_2\}}\\ &\quad-(\mathbf{e}'_{j_2}\mathbf{P}_V\mathbf{e}_{\ell_2})(\mathbf{e}'_{j_1}\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1})|.\end{aligned}\tag{54}$$

Upon defining $x_{j_1,\ell_1}:=\mathbf{e}'_{j_1}\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{\ell_1}$ and $y_{j_2,\ell_2}:=(\mathbf{e}'_{j_2}\mathbf{P}_V\mathbf{e}_{\ell_2})$, squaring $g$ gives rise to

$$\begin{aligned}g^2(j_1,j_2,\ell_1,\ell_2) &= x^2_{j_1,\ell_1}\mathbb{1}_{\{j_2=\ell_2\}}+y^2_{j_2,\ell_2}x^2_{j_1,\ell_1}\\ &\quad-2y_{j_2,\ell_2}x^2_{j_1,\ell_1}\mathbb{1}_{\{j_2=\ell_2\}}.\end{aligned}\tag{55}$$

Since $y_{j_2,\ell_2}\mathbb{1}_{\{j_2=\ell_2\}}=\|\mathbf{P}_V\mathbf{e}_{j_2}\|^2\mathbb{1}_{\{j_2=\ell_2\}}\geq 0$, one can ignore the third summand in (55) to arrive at

$$g(j_1,j_2,\ell_1,\ell_2)\leq x_{j_1,\ell_1}[\mathbb{1}_{\{j_2=\ell_2\}}+y^2_{j_2,\ell_2}]^{1/2}.\tag{56}$$

Toward bounding the scalars $x_{j_1,\ell_1}$ and $y_{j_2,\ell_2}$, rewrite $x_{j_1,\ell_1}:=\mathbf{e}'_{j_1}\mathbf{R}'\mathbf{R}\mathbf{e}_{\ell_1}-\mathbf{e}'_{j_1}\mathbf{R}'\mathbf{P}_U\mathbf{R}\mathbf{e}_{\ell_1}$. If $j_1=\ell_1$, it holds that $x_{j_1,\ell_1}\leq\|\mathbf{R}\mathbf{e}_{\ell_1}\|^2\leq c(1+\delta_1(\mathbf{R}))$; otherwise,

$$\begin{aligned}x_{j_1,\ell_1} &\leq |\mathbf{e}'_{j_1}\mathbf{R}'\mathbf{R}\mathbf{e}_{\ell_1}|+|\mathbf{e}'_{j_1}\mathbf{R}'\mathbf{P}_U\mathbf{R}\mathbf{e}_{\ell_1}|\\ &\leq c\theta_{1,1}(\mathbf{R})+c(1+\delta_1(\mathbf{R}))\gamma^2_R(\mathbf{U}).\end{aligned}$$

Moreover, $y_{j_2,\ell_2}\leq\|\mathbf{P}_V\mathbf{e}_{j_2}\|\|\mathbf{P}_V\mathbf{e}_{\ell_2}\|\leq\gamma^2(\mathbf{V})$. Plugging the bounds into (56) yields

$$\begin{aligned}g(j_1,j_2,\ell_1,\ell_2) &\leq \left[c(1+\delta_1(\mathbf{R}))\mathbb{1}_{\{j_1=\ell_1\}}+c\theta_{1,1}(\mathbf{R})\right.\\ &\quad\left.+c(1+\delta_1(\mathbf{R}))\gamma^2_R(\mathbf{U}))\mathbb{1}_{\{j_1\neq\ell_1\}}\right]\\ &\quad\times[\mathbb{1}_{\{j_2=\ell_2\}}+\gamma^4(\mathbf{V})]^{1/2}.\end{aligned}\tag{57}$$

Plugging (57) into (53), one arrives at

$$\begin{aligned}\|\mathbf{A}_{\Omega^\perp}\mathbf{A}'_\Omega\|_{\infty,\infty} &\leq c[\sqrt{2}k+s\gamma^2(\mathbf{V})]\theta_{1,1}(\mathbf{R})\\ &\quad+c(1+\delta_1(\mathbf{R}))\Big[k\gamma^2(\mathbf{V})+\sqrt{2}k\gamma^2_R(\mathbf{U})\\ &\quad+s\gamma^2_R(\mathbf{U})\gamma^2(\mathbf{V})\Big]\\ &:= c\omega\end{aligned}\tag{58}$$

after using: i) $\mathcal{S}\cap\bar{\mathcal{S}}=\emptyset$ and consequently $j_2\neq\ell_2$ when $j_1=\ell_1$; and ii) $\gamma(\mathbf{V})\leq 1$.

Moving on, consider bounding $\|\mathbf{I}-\mathbf{A}_\Omega\mathbf{A}'_\Omega\|_{\infty,\infty}$ that can be rewritten as

$$\begin{aligned}\|\mathbf{I}-\mathbf{A}_\Omega{\mathbf{A}_\Omega}'\|_{\infty,\infty} &= \max_i\|{\mathbf{e}_i}'(\mathbf{I}-\mathbf{A}_\Omega{\mathbf{A}_\Omega}')\|_1\\ &= \max_i\Big\{|1-\|{\mathbf{e}_i}'\mathbf{A}_\Omega\|^2|\\ &\quad+\sum_{k\neq i}|\langle{\mathbf{e}_i}'\mathbf{A}_\Omega,{\mathbf{e}_k}'\mathbf{A}_\Omega\rangle|\Big\}\\ &= \max_{\substack{j=j_1+j_2\\(j_1,j_2)\in\mathcal{S}}}\Big\{|1-\|\mathbf{A}'\mathbf{e}_j\|^2|\\ &\quad+\sum_{\ell\neq j}|\langle\mathbf{A}'\mathbf{e}_j,\mathbf{A}'\mathbf{e}_\ell\rangle|\Big\}.\end{aligned}\tag{59}$$

In the sequel, an upper bound is derived for (59). Let $(j_1,j_2)$ denote the element of $\mathcal{S}$ associated with $j$ in (59). For the first summand inside the curly brackets in (59), consider lower bounding the norm of the $j$th row of $\mathbf{A}$ as

$$\begin{aligned}\|\mathbf{A}'\mathbf{e}_j\| &= \|(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2}(\mathbf{I}-\mathbf{P}_V)\|_F\\ &= \|\mathcal{P}_{\Phi^\perp}(\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2})\|_F\\ &= \|\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2}-\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2})\|_F\\ &\geq \|\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2}\|-\|\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2})\|_F\\ &\geq \|\mathbf{R}\mathbf{e}_{j_1}\mathbf{e}'_{j_2}\|(1-\mu(\Phi,\Omega_R))\\ &\geq c^{1/2}(1-\delta_1(\mathbf{R}))^{1/2}(1-\mu(\Phi,\Omega_R)).\end{aligned}$$

Since $\delta_1(\mathbf{R})<1$ and $\mu(\Phi,\Omega_R)<1$, one obtains $|1-\|\mathbf{A}'\mathbf{e}_j\|^2|\leq 1-c(1-\delta_1(\mathbf{R}))(1-\mu(\Phi,\Omega_R))^2$.

For the second summand inside the curly brackets in (59), a procedure similar to the one used for bounding $\|\mathbf{A}_{\Omega^\perp}{\mathbf{A}_\Omega}'\|_{\infty,\infty}$ is pursued. First, observe that

$$\begin{aligned}\sum_{\ell\neq j}|\langle\mathbf{A}\mathbf{A}'\mathbf{e}_j,\mathbf{e}_\ell\rangle| &= \sum_{\ell\neq j}|\langle(\mathbf{I}-\mathbf{P}_V)\otimes\mathbf{R}'(\mathbf{I}-\mathbf{P}_U)\mathbf{R}\mathbf{e}_j,\mathbf{e}_\ell\rangle|\\ &= \sum_{(\ell_1,\ell_2)\in\mathcal{S}\setminus\{(j_1,j_2)\}}g(j_1,j_2,\ell_1,\ell_2)\end{aligned}\tag{60}$$

to deduce that, up to a summand corresponding to the index pair $(j_1, j_2)$, (60) is identical to the summation in (53). Following similar arguments to those leading to (57), one arrives at

$$\max_{\substack{j=j_1+j_2\\(j_1,j_2)\in\mathcal{S}}} \sum_{\ell\neq j} |\langle \mathbf{A}'\mathbf{e}_j, \mathbf{A}'\mathbf{e}_\ell\rangle| \leq c\omega.$$

Putting all the pieces together, (59) is bounded as

$$\|\mathbf{I} - \mathbf{A}_\Omega\mathbf{A}'_\Omega\|_{\infty,\infty} \leq 1 - c(1-\delta_1(\mathbf{R}))(1-\mu(\Phi,\Omega_R))^2 + c\omega. \tag{61}$$

Note that because of the assumption $\omega < (1-\delta_1(\mathbf{R}))(1-\mu(\Phi,\Omega_R))^2$, $\|\mathbf{I} - \mathbf{A}_\Omega\mathbf{A}'_\Omega\|_{\infty,\infty} < 1$ as supposed at the beginning of the proof. Substituting (58) and (61) into (52) yields the desired bound. ■

## D. *Proof of Lemma 6*

The proof bears some resemblance with those available for the matrix completion problem [13], and PCP [11]. However, presence of the compression matrix $\mathbf{R}$ gives rise to unique challenges in some stages of the proof, which necessitate special treatment. In what follows, emphasis is placed on the distinct arguments required by the setting here.

The main idea is to obtain first an upper bound on the norm of the linear operator $\pi^{-1}\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi - \mathcal{P}_\Phi$, which is then utilized to upper bound $\mu(\Phi,\Omega_R) = \|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\|$. The former is established in the next lemma; see Appendix E for a proof.

*Lemma 10:* Suppose $\mathcal{S} := \text{supp}(\mathbf{A}_0)$ is drawn according to the Bernoulli model with parameter $\pi$. Let $\Lambda := \sqrt{c(1+\delta_1(\mathbf{R}))[\gamma_R^2(\mathbf{U}) + \gamma^2(\mathbf{V})]}$, and $n := \max\{L, F\}$. Then, there are positive numerical constants $C$ and $\tau$ such that

$$\pi^{-1}\|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi - \pi\mathcal{P}_\Phi\| \leq C\sqrt{\frac{\log(LF)}{\pi}} + \tau\Lambda\log(n) \tag{62}$$

holds with probability higher than $1 - \mathcal{O}\left(n^{-C\pi\Lambda\tau}\right)$, provided that the right-hand side is less than one.

Building on (62), it follows that

$$\begin{aligned}\|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi\| - \pi &\overset{(a)}{\leq} \|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi\| - \pi\|\mathcal{P}_\Phi\|\\ &\overset{(b)}{\leq} \|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi - \pi\mathcal{P}_\Phi\|\\ &\leq C\sqrt{\pi\log(LF)} + \tau\pi\Lambda\log(n)\end{aligned} \tag{63}$$

where (a) and (b) come from $\|\mathcal{P}_\Phi\| \leq 1$ and the triangle inequality, respectively. In addition,

$$\begin{aligned}\|\mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}))\|_F^2 &= |\langle\mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X})), \mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}))\rangle|\\ &= |\langle\mathcal{P}_\Phi(\mathbf{R}(\mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X})))), \mathbf{X}\rangle|\\ &\leq \|\mathcal{P}_\Phi(\mathbf{R}(\mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}))))\|_F\|\mathbf{X}\|_F\end{aligned} \tag{64}$$

for all $\mathbf{X} \in \mathbb{R}^{L\times F}$. Recalling the definition of the operator norm, it follows from (64) that $\mu(\Phi,\Omega_R) \leq \sqrt{c^{-1}(1-\delta_k(\mathbf{R}))^{-1}}\|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi\|^{1/2}$. Plugging the bound (63), the result follows readily. ■

## E. *Proof of Lemma 10*

Start by noting that

$$\begin{aligned}\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}) &= \sum_{i,j}\langle\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}), \mathbf{e}_i\mathbf{e}'_j\rangle\mathbf{e}_i\mathbf{e}'_j\\ &= \sum_{i,j}\langle\mathbf{X}, \mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\rangle\mathbf{e}_i\mathbf{e}'_j\end{aligned}$$

and apply the sampling operator to obtain

$$\mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X})) = \sum_{i,j} b_{i,j}\langle\mathbf{X}, \mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\rangle\mathbf{e}_i\mathbf{e}'_j$$

where $\{b_{i,j}\}$ are Bernoulli-distributed i.i.d. random variables with $\Pr(b_{i,j} = 1) = \pi$. Then,

$$\mathcal{P}_\Phi(\mathbf{R}\mathcal{P}_\Omega(\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}))) = \sum_{i,j} b_{i,j}\langle\mathbf{X}, \mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\rangle\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j). \tag{65}$$

Moreover, since $\mathbf{R}\mathbf{R}' = \mathbf{I}_L$ one finally arrives at

$$\begin{aligned}\mathcal{P}_\Phi(\mathbf{X}) &= \mathcal{P}_\Phi(\mathbf{R}\mathbf{R}'\mathcal{P}_\Phi(\mathbf{X}))\\ &= \sum_{i,j}\langle\mathbf{X}, \mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\rangle\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j).\end{aligned} \tag{66}$$

The next bound will also be useful later on

$$\begin{aligned}\|\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}')\|_F^2 &= \langle\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'), \mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\rangle\\ &= \langle\mathbf{P}_U\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}' + \mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\mathbf{P}_V\\ &\quad - \mathbf{P}_U\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\mathbf{P}_V, \mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\rangle\\ &\overset{(a)}{=} \|\mathbf{P}_U\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\|_F^2 + \|\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\mathbf{P}_V\|_F^2\\ &\quad - \|\mathbf{P}_U\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\|_F^2\|\mathbf{P}_V\mathbf{e}_j\|_F^2\\ &\leq c(1+\delta_1(\mathbf{R}))\gamma_R^2(\mathbf{U})\\ &\quad + c(1+\delta_1(\mathbf{R}))\gamma^2(\mathbf{V}) = \Lambda^2\end{aligned} \tag{67}$$

where (a) holds because $\langle\mathbf{P}_U\mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\mathbf{P}_V, \mathbf{R}\mathbf{e}_i{\mathbf{e}_j}'\rangle = \langle\mathbf{e}'_i\mathbf{R}\mathbf{P}_U\mathbf{R}\mathbf{e}_i, \mathbf{e}'_j\mathbf{P}_V\mathbf{e}_j\rangle$ and $\mathbf{P}_U = \mathbf{P}_U^2$ (likewise $\mathbf{P}_V$).

Defining the random variable $\Xi := \pi^{-1}\|\mathcal{P}_\Phi\mathbf{R}\mathcal{P}_\Omega\mathbf{R}'\mathcal{P}_\Phi - \pi\mathcal{P}_\Phi\|$ and using (66), one can write

$$\begin{aligned}\Xi &= \pi^{-1}\sup_{\|\mathbf{X}\|_F=1}\left\|\sum_{i,j}(b_{i,j}-\pi)\langle\mathbf{X}, \mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\rangle\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\right\|_F\\ &= \pi^{-1}\sup_{\|\text{vec}(\mathbf{X})\|=1}\left\|\sum_{i,j}(b_{i,j}-\pi)\text{vec}(\mathbf{X})'\text{vec}[\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)]\right.\\ &\qquad\left.\otimes\,\text{vec}[\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)]\right\|\\ &= \pi^{-1}\left\|\sum_{i,j}(b_{i,j}-\pi)\text{vec}[\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)]\otimes\text{vec}[\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)]\right\|.\end{aligned} \tag{68}$$

Random variables $\{b_{i,j} - \pi\}$ are i.i.d. with zero mean, and thus, one can utilize the spectral concentration inequality in [45, Lemma 3.5] to find

$$\mathbb{E}[\Xi] \leq C\sqrt{\frac{\log(LF)}{\pi}}\max_{i,j}\|\mathcal{P}_\Phi(\mathbf{R}\mathbf{e}_i\mathbf{e}'_j)\|_F \overset{(b)}{\leq} C\sqrt{\frac{\log(LF)}{\pi}}\Lambda \tag{69}$$

for some constant $C > 0$, where $(\mathrm{b})$ is due to (67). Now, applying Talagrand's concentration tail bound [47] to the random variable $\Xi$ yields

$$\Pr(|\Xi - \mathbb{E}[\Xi]| \geq t) \leq 3\exp\left(-\frac{t\log(2)}{K}\pi\min\{1,t\}\right) \quad (70)$$

for some constant $K > 0$, where $t := \tau\Lambda\log(n)$ and $n := \max\{L, F\}$. The arguments leading to (69) and (70) are similar those used in [13, Th. 4.2] for the matrix completion problem, and details are omitted here. Putting (69) and (70) together, it is possible to infer

$$\Xi \leq \mathbb{E}[\Xi] + t \leq C\sqrt{\frac{\log(LF)}{\pi}} + \tau\Lambda\log(n) \quad (71)$$

with probability higher than $1 - \mathcal{O}(n^{-C\pi\Lambda\tau})$, which completes the proof of the lemma. ■

## References

[1] [Online]. Available: http://internet2.edu/observatory/archive/data-collections.html

[2] A. Agarwal, S. Negahban, and M. J. Wainright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," *Ann. Statist.*, vol. 40, pp. 1171–1197, Sep. 2012.

[3] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[4] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, pp. 183–202, Jan. 2009.

[5] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena-Scientific, 1999.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 2nd ed. Belmont, MA, USA: Athena-Scientific, 1999.

[7] P. J. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, pp. 1705–1732, Apr. 2009.

[8] B. Bollobas, *Random Graphs*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, pp. 1–122, 2011.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[11] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 1, pp. 1–37, 2011.

[12] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2009.

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

[14] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[15] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[16] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 14–20, Mar. 2008.

[17] V. Chandrasekaran, S. Sanghavi, P. R. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.

[18] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Trans. Inf. Theory*, 2013, (see also arXiv:1104.0354v2 [cs.IT]).

[19] Q. Chenlu and N. Vaswani, "Recursive sparse recovery in large but correlated noise," in *Proc. 49th Allerton Conf. Commun., Control, Comput.*, Sep. 2011, pp. 752–759.

[20] A. Chistov and D. Grigorev, "Complexity of quantifier elimination in the theory of algebraically closed fields," in *Math. Found. of Computer Science*. Berlin, Germany: Springer-Verlag, 1984, vol. 176, Lecture Notes in Computer Science, pp. 17–31.

[21] F. Deutsch, *Best Approximation in Inner Product Spaces*, 2nd ed. New York, NY, USA: Springer-Verlag, 2001.

[22] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, Dec. 2011.

[23] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell_1$ minimization," *Proc. Natl. Acad. Sci.*, vol. 100, pp. 2197–2202, Mar. 2003.

[24] J. Finn, K. Nael, V. Deshpande, O. Ratib, and G. Laub, "Cardiac MR imaging: State of the technology," *Radiology*, vol. 241, no. 2, pp. 338–354, 2006.

[25] H. Gao, "Prior rank, intensity and sparsity model (PRISM): A divide-and-conquer matrix decomposition model with low-rank coherence and sparse variation," presented at the SPIE Opt. Eng. Appl., 2012.

[26] H. Gao, J. Cai, Z. Shen, and H. Zhao, "Robust principal component analysis-based four-dimensional computed tomography," *Phys. Med. Biol.*, vol. 56, no. 11, pp. 3181–3198, 2011.

[27] B. He and X. Yuan, On the $\mathrm{o}(1/\mathrm{t})$ convergence rate of alternating direction method Nanjing Univ., Nanjing, China, Tech. Rep., 2011.

[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[29] P. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 57–60.

[30] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1025–1031, Sep. 2011.

[31] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer, 2002.

[32] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proc. ACM SIGCOMM*, Portland, OR, USA, Aug. 2004, pp. 219–230.

[33] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, May 2007.

[34] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix Univ. Illinois at Urbana–Champaign, Urbana, USA, 2009, Tech. Rep. UILU-ENG-09-2214.

[35] B. Madore *et al.*, "Unaliasing by Fourier-encoding the overlaps using the temporal dimension (UNFOLD), applied to cardiac imaging and fMRI," *Magn. Reson. Med.*, vol. 42, no. 5, pp. 813–828, 1999.

[36] M. Mardani, G. Mateos, and G. B. Giannakis, "Unveiling anomalies in large-scale networks via sparsity and low rank," in *Proc. 45th Asilomar Conf. Signal, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2011, pp. 403–407.

[37] M. Mardani, G. Mateos, and G. B. Giannakis, "In-network sparsity-regularized rank minimization: Applications and algorithms," *IEEE Trans. Signal Process.*, 2013, (see also arXiv:1203.1507v1 [cs.MA]).

[38] M. Mardani, G. Mateos, and G. B. Giannakis, "Dynamic anomalography: Tracking network anomalies via sparsity and low rank," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 50–66, Feb. 2013.

[39] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[40] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.

[41] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $o(1/k^2)$," *Soviet Math. Doklady*, vol. 27, pp. 372–376, 1983.

[42] Y. Nesterov, "Smooth minimization of nonsmooth functions," *Math. Prog.*, vol. 103, pp. 127–152, 2005.

[43] H. Rauhut, "Compressive sensing and structured random matrices," *Theor. Found. Numer. Methods Sparse Recov.*, vol. 9, pp. 1–92, 2010.

[44] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 662–676, Jun. 2012.

[45] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *J. ACM*, vol. 54, pp. 1–20, Dec. 2006.

[46] P. Sprechmann, A. M. Bronstein, and G. Sapiro, “Real-time online singing voice separation from monaural recordings using robust low-rank modeling,” presented at the Annu. Conf. Int. Soc. Music Inf. Retriev., Porto, Portugal, Oct. 2012.

[47] M. Talagrand, “New concentration inequalities in product spaces,” *Invent. Math.*, vol. 126, pp. 505–563, Dec. 1996.

[48] M. Thottan and C. Ji, “Anomaly detection in IP networks,” *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.

[49] K.-C. Toh and S. Yun, “An accelerated proximal gradient algorithm for nuclear norm regularized least-squares problems,” *Pacific J. Opt.*, vol. 6, pp. 615–640, 2010.

[50] A. E. Waters, A. C. Sankaranarayanan, and R. G. Baraniuk, “SpaRCS: Recovering low-rank and sparse matrices from compressive measurements,” presented at the Proc. Neural Inf. Process. Syst., Granada, Spain, Dec. 2011.

[51] J. Wright, A. Ganesh, and K. M. Y. Ma, “Compressive principal component pursuit,” in *Proc. Int. Symp. Inf. Theory*, Cambridge, MA, USA, Jul. 2012, pp. 1276–1280.

[52] L. Xing *et al.*, “Overview of image-guided radiation therapy,” *Med. Dosimetry*, vol. 31, no. 2, pp. 91–112, 2006.

[53] H. Xu, C. Caramanis, and S. Sanghavi, Robust PCA via Outlier Pursuit 2010 [Online]. Available: arXiv:1010.4237v2 [cs.LG]

[54] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, “Network anomography,” in *Proc. Internet Meas. Conf.*, Berkeley, CA, USA, Oct. 2005, pp. 317–330.

[55] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma, “Stable principal component pursuit,” in *Proc. Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 2010, pp. 1518–1522.

**Morteza Mardani** (S’06) received his B.Sc. degree in Electrical Engineering from the Shahid Bahonar University of Kerman, Kerman, Iran, in 2006, and the M.Sc. degree in Electrical Engineering from the University of Tehran, Tehran, Iran, in 2009. Since September 2009, he has been working toward his Ph.D. degree with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis. His research interests include network science, statistical learning, and compressive sampling.

Mr. Mardani is the recipient of the Best Student Paper Award from the 13th IEEE Workshop on Signal Processing Advances in Wireless Communications. He also received the ADC Fellowship Award from the Digital Technology Center at the University of Minnesota for two academic years 2009–2010 and 2010–2011.

**Gonzalo Mateos** (M’12) received his B.Sc. degree in Electrical Engineering from Universidad de la República (UdelaR), Montevideo, Uruguay in 2005 and the M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from the University of Minnesota, Minneapolis, in 2009 and 2011.

Since 2012, he has been a post doctoral research associate with the Department of Electrical and Computer Engineering and the Digital Technology Center, University of Minnesota. Since 2003, he is an assistant with the Department of Electrical Engineering, UdelaR. From 2004 to 2006, he worked as a Systems Engineer at Asea Brown Boveri (ABB), Uruguay. His research interests lie in the areas of communication theory, signal processing and networking. His current research focuses on distributed signal processing, sparse linear regression, and statistical learning for social data analysis and network health monitoring.

**Georgios B. Giannakis** (F’97) received his Diploma in Electrical Engineering from the National Technical University of Athens, Greece, 1981. From 1982 to 1986 he was with the University of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engineering, 1986. Since 1999 he has been a professor with the University of Minnesota, where he now holds an ADC Chair in Wireless Telecommunications in the ECE Department, and serves as director of the Digital Technology Center.

His general interests span the areas of communications, networking and statistical signal processing – subjects on which he has published more than 350 journal papers, 580 conference papers, 20 book chapters, two edited books and two research monographs (h-index 104). Current research focuses on sparsity in signals and systems, wireless cognitive radios, mobile ad hoc networks, renewable energy, power grid, gene-regulatory, and social networks. He is the (co-) inventor of 21 patents issued, and the (co-) recipient of 8 best paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.