Tight convex relaxations for sparse matrix factorization

Emile Richard, Guillaume Obozinski, Jean-Philippe Vert

Tight convex relaxations for sparse matrix factorization

Emile Richard[1], Guillaume Obozinski[2] and Jean-Philippe Vert[3,4,5]

[1]Department of Electrical Engineering, Stanford University
[2]Université Paris-Est, Laboratoire d'Informatique Gaspard Monge,
Groupe Imagine, Ecole des Ponts - ParisTech, 77455 Marne-la-Vallée, France
[3]Mines ParisTech, PSL Research University,
CBIO-Centre for Computational Biology, 77300 Fontainebleau, France
[4]Institut Curie, 75248 Paris Cedex ,France
[5]INSERM U900, 75248 Paris Cedex ,France

Abstract

Based on a new atomic norm, we propose a new convex formulation for sparse matrix factorization problems in which the number of nonzero elements of the factors is assumed fixed and known. The formulation counts sparse PCA with multiple factors, subspace clustering and low-rank sparse bilinear regression as potential applications. We compute slow rates and an upper bound on the statistical dimension Amelunxen et al. (2013) of the suggested norm for rank 1 matrices, showing that its statistical dimension is an order of magnitude smaller than the usual ℓ_1-norm, trace norm and their combinations. Even though our convex formulation is in theory hard and does not lead to provably polynomial time algorithmic schemes, we propose an active set algorithm leveraging the structure of the convex problem to solve it and show promising numerical results.

1 Introduction

A range of machine learning problems such as link prediction in graphs containing community structure (Richard et al., 2014), phase retrieval (Candès et al., 2013), subspace clustering (Wang et al., 2013) or dictionary learning for sparse coding (Mairal et al., 2010) amount to solve sparse matrix factorization problems, *i.e.*, to infer a low-rank matrix that can be factorized as the product of two sparse matrices with few columns (left factor) and few rows (right factor). Such a factorization allows for more efficient storage, faster computation, more interpretable solutions, and, last but not least, it leads to more accurate estimates in many situations. In the case of interaction networks for example, the assumption that the network is organized as a collection of highly connected communities which can overlap implies that the adjacency matrix admits such a factorization. More generally, considering sparse low-rank matrices combines two natural forms of sparsity, in the spectrum and in the support, which can be motivated by the need to explain systems behaviors by a superposition of latent processes which only involve a few parameters. Landmark applications of sparse matrix factorization are sparse principal components analysis (SPCA, d'Aspremont et al., 2007; Zou et al., 2006) or sparse canonical correlation analysis (SCCA, Witten et al., 2009), which are widely used to analyze high-dimensional data such as genomic data.

From a computational point of view, however, sparse matrix factorization is challenging since it typically leads to non-convex, NP-hard problems (Moghaddam et al., 2006). For instance, Berthet and Rigollet (2013) noted that solving sparse PCA with a single component is equivalent to the planted clique

problem (Jerrum, 1992), a notoriously hard problem when the size of the support is smaller than the square root of size of the matrix. Many heuristics and relaxations have therefore been proposed, with and without theoretical guaranties, to approximatively solve the problems leading to sparse low-rank matrices. A popular procedure is to alternatively optimize over the left and right factors in the factorization, formulating each step as a convex optimization problem (Lee et al., 2007; Mairal et al., 2010). Despite these worst case computational hardness, simple generalizations of the power method have been proposed by Journée et al. (2010); Luss and Teboulle (2013); Yuan and Zhang (2013) for the sparse PCA problem with a single component. These algorithms perform well empirically and have been proved to be efficient theoretically under mild conditions by Yuan and Zhang (2013). Several semidefinite programming (SDP) convex relaxations of the same problem have also been proposed (Amini and Wainwright, 2009; d'Aspremont et al., 2007, 2008). Based on the rank one approximate solutions, computing multiple principal components of the data is commonly done though successive deflations (Mackey, 2009) of the input matrix.

Recently, several authors have investigated the possibility to formulate sparse matrix factorization as a convex optimization problem. Bach et al. (2008) showed that the convex relaxation of a number of natural sparse factorization are too coarse too succeed, while Bach (2013) investigated several convex formulations involving nuclear norms (Jameson, 1987), similar to the ones we investigate in this paper, and their SDP relaxations. Several authors also investigated the performance of regularizing a convex loss with linear combinations of the ℓ_1 norm and the trace norm, naturally leading to a matrix which is both sparse and low-rank (Doan and Vavasis, 2013; Oymak et al., 2012; Richard et al., 2012, 2013, 2014). This penalty term can be related to the SDP relaxations of d'Aspremont et al. (2007, 2008) that penalize the trace and the element-wise ℓ_1 norm of the positive semi-definite unknown. The statistical performance of these basic combinations of the two convex criteria has however been questioned by Krauthgamer et al. (2013); Oymak et al. (2012). Oymak et al. (2012) showed that for compressed sensing applications, no convex combination of the two norms improves over each norm taken alone. Krauthgamer et al. (2013) prove that the SDP relaxations fail at finding the sparse principal component outside the favorable regime where a simple diagonal thresholding algorithm (Amini and Wainwright, 2009) works. Moreover, these existing convex formulations either aim at finding only a rank one matrix, or a low rank matrix whose factors themselves are not necessarily guaranteed to be sparse.

In this work, we propose two new matrix norms which, when used as regularizer for various optimization problems, do yield estimates for low-rank matrices with multiple sparse factors that are provably more efficient statistically than the ℓ_1 and trace norms. The price to pay for this statistical efficiency is that, although convex, the resulting optimization problems are NP-hard, and we must resort to heuristic procedures to solve them. Our numerical experiments however confirm that we obtain the desired theoretical gain to estimate low-rank sparse matrices.

1.1 Contributions and organization of the paper

More precisely, our contributions are:

- **Two new matrix norms (Section 2).** In order to properly define matrix factorization, given sparsity levels of the factors denoted by integers k and q, we first introduce in Section 2.1 the (k, q)-rank of a matrix as the minimum number of left and right factors, having respectively k and q nonzeros, required to reconstruct a matrix. This index is a more involved complexity measure for matrices than the rank in that it conditions on the number of nonzero elements of the left and right factors of a matrix. Using this index, we propose in Section 2.2 two new *atomic norms* for matrices (Chandrasekaran et al., 2012). (i) Considering the convex hull unit

operator norm matrices with (k,q)-rank $= 1$, we build a convex surrogate to low (k,q)-rank matrix estimation problem. (*ii*) We introduce a polyhedral norm built upon (k,q)-rank $= 1$ matrices with all non-zero entries of absolute value equal to 1. We provide in Section 2.3 an equivalent characterization of the norms as nuclear norms, in the sense of Jameson (1987), highlighting in particular a link to the k-support norm of Argyriou et al. (2012).

- **Using these norms to estimate sparse low-rank matrices (Section 3).** We show how several problems such as bilinear regression or sparse PCA can be formulated as convex optimization problems with our new norms, and clarify that the resulting problems can however be NP-hard.

- **Statistical Analysis (Section 4).** We study the statistical performance of the new norms and compare them with existing penalties. Our analysis goes first in Section 4.1 using *slow rate* type of upper bounds on the denoising error, which despite sub-optimality gives a first insight on the gap between the statistical performance of our (k,q)-trace norm and that of the ℓ_1 and trace norms. Next we show in Section 4.2, using cone inclusions and estimates of statistical dimension, that our norms are superior to any convex combination of the trace norm and the ℓ_1 norm in a number of different tasks. However, our analysis also shows that the factors gained over the rivals to estimate sparse low-rank matrices vanishes when we use our norm to estimate sparse vectors.

- **A working set algorithm (Section 5).** While in the vector case the computation remains feasible in polynomial time, the norms we introduce for matrices can not be evaluated in polynomial time. We propose algorithmic schemes to approximately learn with the new norms. The same norms and meta-algorithms can be used as a regularizer in supervised problems such as bilinear and quadratic regression. Our algorithmic contribution does not consist in providing more efficient solutions to the rank-1 SPCA problem, but to combine atoms found by the rank-1 solvers in a principled way.

- **Numerical experiments (Section 6).** We numerically evaluate the performance of our new norms on simulated data, and confirm the theoretical results. While our theoretical analysis only focuses on the estimation of sparse matrices with (k,q)-rank one, our simulations allow us to conjecture that the statistical dimension scales linearly with the (k,q)-rank and decays with the overlap between blocks. We also show that our model is competitive with the state-of-the-art on the problem of sparse PCA.

Due to their length and technicality, all proofs are postponed to the appendices.

1.2 Notations

For any integers $1 \leq k \leq p$, $[1,p] = \{1,\dots,p\}$ is the set of integers from 1 to p and $\mathcal{G}_k^p$ denotes the set of subsets of k indices in $[1,p]$. For a vector $w \in \mathbb{R}^p$, $\|w\|_0$ is the number of non-zero coefficients in w, $\|w\|_1 = \sum_{i=1}^p |w_i|$ is its ℓ_1 norm, $\|w\|_2 = \left(\sum_{i=1}^p w_i^2\right)^{\frac{1}{2}}$ is its Euclidean norm, $\|w\|_\infty = \max_i |w_i|$ is its ℓ_∞ norm and $\mathrm{supp}(w) \in \mathcal{G}^p_{\|w\|_0}$ is its support, *i.e.*, the set of indices of the nonzero entries of w. For any $I \subset [1,p]$, $w_I \in \mathbb{R}^p$ is the vector that is equal to w on I, and has 0 entries elsewhere. Given matrices A and B of the same size, $\langle A,B\rangle = \mathrm{tr}(A^\top B)$ is the standard inner product of matrices. For any matrix $Z \in \mathbb{R}^{m_1 \times m_2}$ the notations $\|Z\|_0$, $\|Z\|_1$, $\|Z\|_\infty$, $\|Z\|_{\mathrm{Fro}}$, $\|Z\|_*$, $\|Z\|_{\mathrm{op}}$ and $\mathrm{rank}(Z)$ stand respectively for the number of nonzeros, entry-wise ℓ_1 and ℓ_∞ norms, the standard ℓ_2 (or Frobenius) norm, the trace-norm (or nuclear norm, the sum of the singular values), the operator

norm (the largest singular value) and the rank of Z, while $\mathrm{supp}(Z) \subset [1, m_1] \times [1, m_2]$ is the support of Z, *i.e.*, the set of indices of nonzero elements of Z. When dealing with a matrix Z whose nonzero elements form a block of size $k \times q$, $\mathrm{supp}(Z)$ takes the form $I \times J$ where $(I, J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}$. For a matrix Z and two subsets of indices $I \subset [1, m_1]$ and $J \subset [1, m_2]$, $Z_{I,J}$ is the matrix having the same entries as Z inside the index subset $I \times J$, and 0 entries outside. This notation should not be confused with the notation $Z^{(IJ)}$ which we will sometimes use to denote a general matrix with support contained in $I \times J$.

2 Tight convex relaxations of sparse factorization constraints

In this section we propose two new matrix norms allowing to formulate various sparse matrix factorization problems as convex optimization problems. We start by defining the (k, q)-rank of a matrix in Section 2.1, a useful generalization of the rank which also quantifies the sparseness of a matrix factorization. We then introduce two atomic norms defined as tight convex relaxations of the (k, q)-rank in Section 2.2: the (k, q)-trace norm, obtained by relaxing the (k, q)-rank over the operator norm ball, and the (k, q)-CUT norm, obtained by a similar construction with extra-constraints on the element-wise ℓ_∞ of factors. In Section 2.3 we relate these matrix norms to vector norms using the concept of nuclear norms, establishing in particular a connection of the (k, q)-trace norm for matrices with the k-support norm of Argyriou et al. (2012), and the (k, q)-CUT norm to the vector k-norm, defined as the sum of the k largest components in absolute value of a vector (Bhatia, 1997, Exercise II.1.15).

2.1 The (k, q)-rank of a matrix

The rank of a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$ is the minimum number of rank-1 matrices (*i.e.*, outer products of vectors of the form $ab^\top$ for $a \in \mathbb{R}^{m_1}$ and $b \in \mathbb{R}^{m_2}$) needed to express Z as a linear combination of the form $Z = \sum_{i=1}^r a_i b_i^\top$. It is a versatile concept in linear algebra, central in particular to solve matrix factorization problems and low-rank approximations. The following definition generalizes this notion to incorporate constraints on the sparseness of the rank-1 elements:

Definition 1 ((k, q)-SVD and (k, q)-rank) *For a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, we call (k, q)-sparse singular value decomposition of Z (or (k, q)-SVD) any decomposition of the form $Z = \sum_{i=1}^r c_i a_i b_i^\top$ where $c_1 \geq c_2 \geq \cdots \geq c_r > 0$, a_i (resp. b_i) are unit vectors with at most k (resp. q) nonzero elements, and with minimal r, which we call the (k, q)-rank of Z. In such a decomposition, we refer to vectors a_i and b_i as left and right (k, q)-sparse singular vectors of Z, and to c_i as its (k, q)-sparse singular values.*

The (k, q)-rank and (k, q)-SVD of Z can equivalently be defined as the optimal value and one of the solutions of the optimization problem:

$$\min \|c\|_0 \quad \text{s.t.} \quad Z = \sum_{i=1}^{\infty} c_i a_i b_i^\top, \qquad (a_i, b_i, c_i) \in \mathcal{A}_k^{m_1} \times \mathcal{A}_q^{m_2} \times \mathbb{R}_+, \tag{1}$$

where for any $1 \leq j \leq n$,

$$\mathcal{A}_j^n = \{a \in \mathbb{R}^n \ : \ \|a\|_0 \leq j, \|a\|_2 = 1\}.$$

refers to the set of n-dimensional unit vectors with at most j non-zero components. When $k = m_1$ and $q = m_2$, we recover the usual notions of rank and SVD of a matrix. In general, however, the (k, q)-rank and (k, q)-SVD do not share several important properties of the usual rank and SVD, as the following proposition shows:

Proposition 2 *(Properties of the (k,q)-SVD)*

1. *The (k,q)-rank of a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$ can be strictly larger than m_1 and m_2.*
2. *The (k,q)-SVD is not necessarily unique.*
3. *The (k,q)-sparse singular vectors are not necessarily orthogonal to each other.*

For $k = q = 1$, the $(1,1)$-SVD decomposes Z as a sum of matrices with only one non-zero element, showing that $(1,1)\text{-rank}(Z) = \|Z\|_0$. Since $\mathcal{A}_i^n \subset \mathcal{A}_j^n$ when $i \leq j$, we deduce from the expression of the (k,q)-rank as the optimal value of (1) that the following tight inequalities hold:

$$\forall (k,q) \in [1, m_1] \times [1, m_2], \quad \text{rank}(Z) \;\leq\; (k,q)\text{-rank}(Z) \;\leq\; \|Z\|_0 \,.$$

The (k,q)-rank is useful to formulate problems in which a matrix should be modeled as or approximated by a matrix with sparse low rank factors, with the assumption that the sparsity level of the factors is fixed and known. For example, the standard rank-1 SPCA problem consists in finding the symmetric matrix with (k,k)-rank equal to 1 and providing the best approximation of the sample covariance matrix (Zou et al., 2006).

2.2 Two convex relaxations for the (k,q)-rank

The (k,q)-rank is obviously a discrete, nonconvex index, like the rank or the cardinality, leading to computational difficulties when one wants to estimate matrices with small (k,q)-rank. In this section, we propose two convex relaxations of the (k,q)-rank aimed at mitigating these difficulties. They are both instances of the atomic norms introduced by Chandrasekaran et al. (2012), which we first review.

Definition 3 (Atomic norm) *Given a centrally symmetric compact subset $\mathcal{A} \subset \mathbb{R}^p$ of elements called atoms, the* atomic norm *induced by $\mathcal{A}$ on $\mathbb{R}^p$ is the gauge function*[1] *of $\mathcal{A}$, defined by*

$$\|x\|_{\mathcal{A}} = \inf \{t > 0 \;:\; x \in t\, conv(\mathcal{A})\} \,, \tag{2}$$

where $conv(\mathcal{A})$ denotes the convex hull of $\mathcal{A}$.

Chandrasekaran et al. (2012) show that the atomic norm induced by $\mathcal{A}$ is indeed a norm, which can be rewritten as

$$\|x\|_{\mathcal{A}} = \inf \left\{ \sum_{a \in \mathcal{A}} c_a \;:\; x = \sum_{a \in \mathcal{A}} c_a a, \;\; c_a \geq 0, \;\; \forall a \in \mathcal{A} \right\} , \tag{3}$$

and whose dual norm satisfies

$$\begin{aligned} \|x\|_{\mathcal{A}}^* &:= \sup \{\langle x, z \rangle \;:\; \|z\|_{\mathcal{A}} \leq 1\} \\ &= \sup \{\langle x, a \rangle \;:\; a \in \mathcal{A}\} \,. \end{aligned} \tag{4}$$

We can now define our first convex relaxation of the (k,q)-rank:

Definition 4 ((k,q)-trace norm) *For a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, the (k,q)-trace norm $\Omega_{k,q}(Z)$ is the atomic norm induced by the set of atoms:*

$$\mathcal{A}_{k,q} = \left\{ ab^\top \;:\; a \in \mathcal{A}_k^{m_1}, \;\; b \in \mathcal{A}_q^{m_2} \right\} . \tag{5}$$

[1] see Rockafellar (1997), p. 28, for a precise definition of gauge functions.

In words, $\mathcal{A}_{k,q}$ is the set of matrices $Z \in \mathbb{R}^{m_1 \times m_2}$ such that (k,q)-rank$(Z) = 1$ and $\|Z\|_{\text{op}} = 1$. Plugging (5) into (3), we obtain an equivalent definition of the (k,q)-trace norm as the optimal value of the following optimization problem:

$$\Omega_{k,q}(Z) = \min \left\{ \|c\|_1 \ : \ Z = \sum_{i=1}^{\infty} c_i a_i b_i^\top, \ (a_i, b_i, c_i) \in \mathcal{A}_k^{m_1} \times \mathcal{A}_q^{m_2} \times \mathbb{R}_+ \right\} . \tag{6}$$

Comparing (6) to (1) shows that the (k,q)-trace norm is derived from the (k,q)-rank by replacing the non-convex ℓ_0 pseudo-norm of c by its convex ℓ_1 norm in the optimization problem. In particular, in the case $k = m_1$ and $q = m_2$, the (k,q)-trace norm is the usual trace norm (equal to the ℓ_1-norm of singular values), i.e. the usual relaxation of the rank (which is the ℓ_0-norm of the singular values). Similarly, when $k = q = 1$, the (k,q)-trace norm is simply the ℓ_1 norm. Just like the (k,q)-rank interpolates between the ℓ_0 pseudo-norm and the rank, the (k,q)-trace norm interpolates between the ℓ_1 norm and the trace norm. Indeed, since $\mathcal{A}_i^n \subset \mathcal{A}_j^n$ when $i \leq j$, we deduce from the expression of $\Omega_{k,q}$ as the optimal value of (6) that the following tight inequalities hold for any $1 \leq k \leq m_1$ and $1 \leq q \leq m_2$:

$$\Omega_{m_1,m_2}(Z) = \|Z\|_* \leq \Omega_{k,q}(Z) \leq \|Z\|_1 = \Omega_{1,1}(Z) . \tag{7}$$

While the SVD decomposition of a matrix used to define its trace norm is the same as the one used to define its rank, this may not be the case anymore for the (k,q)-trace norm. Indeed, in the general case, the (k,q)-trace norm may not be simply the sum of (k,q)-sparse singular values associated to a (k,q)-SVD according to Definition 1, because the vectors c that solve (6) and (1) can be different. This justifies the following definition:

Definition 5 (Soft-(k,q)-SVD and soft-(k,q)-rank) *For a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, we call soft-$(k,q)$-sparse singular value decomposition (or soft-(k,q)-SVD) any decomposition $Z = \sum_{i=1}^r c_i a_i b_i^\top$ that solves (6) with $c_1 \geq c_2 \geq \ldots \geq c_r > 0$. The soft-$(k,q)$-rank of Z is the minimum number of terms in a soft-(k,q)-SVD of Z.*

Similar to the (k,q)-SVD, the soft-(k,q)-SVD lacks many important properties of the trace norm when $k < m_1$ and $q < m_2$:

Proposition 6 *1. The soft-(k,q)-rank of a matrix can be strictly larger than its (k,q)-rank.*

2. The soft-(k,q)-SVD is not necessarily unique.

3. The soft-(k,q)-sparse singular vectors are not necessarily orthogonal to each other.

In addition to (6), the next lemma provides another explicit formulation for the (k,q)-trace norm, its dual and its sub differential:

Lemma 7 *For any $Z \in \mathbb{R}^{m_1 \times m_2}$ we have*

$$\Omega_{k,q}(Z) = \inf \left\{ \sum_{(I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}} \left\| Z^{(I,J)} \right\|_* \ : \ Z = \sum_{(I,J)} Z^{(I,J)} \ , \ \mathit{supp}(Z^{(I,J)}) \subset I \times J \right\} , \tag{8}$$

and

$$\Omega_{k,q}^*(Z) = \max \left\{ \|Z_{I,J}\|_{\text{op}} \ : \ I \in \mathcal{G}_k^{m_1} \ , \ J \in \mathcal{G}_q^{m_2} \right\} . \tag{9}$$

The subdifferential of $\Omega_{k,q}$ at an atom $A = ab^\top \in \mathcal{A}_{k,q}$ with $I_0 = \text{supp}(a)$ and $J_0 = \text{supp}(b)$ is

$$\partial \Omega_{k,q}(A) = \left\{ A + Z \ : \ A Z_{I_0,J_0}^\top = 0, \ A^\top Z_{I_0,J_0} = 0, \ \forall (I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2} \ \ \|A_{I,J} + Z_{I,J}\|_{\text{op}} \leq 1 \right\} . \tag{10}$$

Our second norm is again an atomic norm, but is obtained by focusing on a more restricted set of atoms. It is motivated by applications where we want to estimate matrices which, in addition to being sparse and low-rank, are constant over blocks, such as adjacency matrices of graphs with non-overlapping communities. For that purpose, consider first the subset of $\mathcal{A}_k^m$ made of vectors whose nonzero entries are all equal in absolute value:

$$\widetilde{\mathcal{A}}_k^m = \left\{ a \in \mathbb{R}^m,\ \|a\|_0 = k\ ,\ \forall i \in \mathrm{supp}(a),\ |a_i| = \tfrac{1}{\sqrt{k}} \right\}.$$

We can then define our second convex relaxation of the (k,q)-rank:

Definition 8 ((k,q)-CUT norm) *We define the (k,q)-CUT norm $\widetilde{\Omega}_{k,q}(Z)$ as the atomic norm induced by the set of atoms*

$$\widetilde{\mathcal{A}}_{k,q} = \left\{ ab^\top \ :\ a \in \widetilde{\mathcal{A}}_k^{m_1},\ b \in \widetilde{\mathcal{A}}_q^{m_2} \right\}. \tag{11}$$

In other words, the atoms in $\widetilde{\mathcal{A}}_{k,q}$ are the atoms of $\mathcal{A}_{k,q}$ whose nonzero elements all have the same amplitude.

Our choice of terminology is motivated by the following relation of our norm to the CUT-polytope: in the case $k = m_1$ and $q = m_2$, the unit ball of $\widetilde{\Omega}_{k,q}$ coincides (up to a scaling factor of $\sqrt{m_1 m_2}$) with the polytope known as the CUT polytope of the complete graph on n vertices (Deza and Laurent, 1997), defined by

$$\mathrm{CUT} = \mathrm{conv}\left\{ ab^\top\ ,\ a \in \{\pm 1\}^{m_1}\ ,\ b \in \{\pm 1\}^{m_2} \right\}.$$

The norm obtained as the gauge of the CUT polytope is therefore to the trace norm as $\widetilde{\Omega}_{k,q}$ is to $\Omega_{k,q}$.

2.3 Equivalent nuclear norms built upon vector norms

In this section we show that the (k,q)-trace norm (Definition 4) and the (k,q)-CUT norm (Definition 8), which we defined as atomic norms induced by specific atom sets, can alternatively be seen as instances of *nuclear norms* considered by Jameson (1987). For that purpose it is useful to recall the general definition of nuclear norms and the characterization of the corresponding dual norms as formulated in Jameson (1987, Propositions 1.9 and 1.11):

Proposition 9 (nuclear norm) *Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ denote any vector norms on $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$, respectively, then*

$$\nu(Z) := \inf\left\{ \sum_i \|a_i\|_\alpha \|b_i\|_\beta \ :\ Z = \sum_i a_i b_i^\top \right\},$$

where the infimum is taken over all summations of finite length, is a norm over $\mathbb{R}^{m_1 \times m_2}$ called the nuclear norm *induced by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$. Its dual is given by*

$$\nu^*(Z) = \sup\left\{ a^\top Z b\ :\ \|a\|_\alpha \le 1\ ,\ \|b\|_\beta \le 1 \right\}. \tag{12}$$

The following lemma shows that the nuclear norm induced by two atomic norms is itself an atomic norm.

Lemma 10 *If $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two atomic norms on $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$ induced respectively by two atom sets $\mathcal{A}_1$ and $\mathcal{A}_2$, then the nuclear norm on $\mathbb{R}^{m_1 \times m_2}$ induced by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ is an atomic norm induced by the atom set:*

$$\mathcal{A} = \left\{ ab^\top\ :\ a \in \mathcal{A}_1\ ,\ b \in \mathcal{A}_2 \right\}.$$

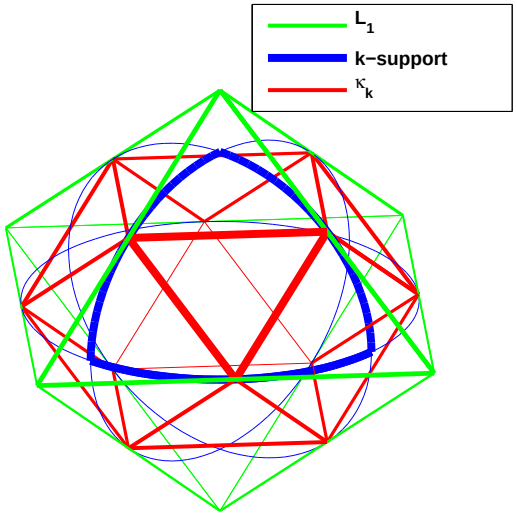

Figure 1: Unit balls of 3 norms of interest for vectors of $\mathbb{R}^3$ materialized by their sets of extreme points at which the norm is non-differentiable. Each unit ball is the convex hull of the corresponding sets. In green, the usual ℓ_1-norm scaled by the factor $1/\sqrt{k} = 1/\sqrt{2}$, in blue the norm θ_2 (a.k.a. 2-support norm), in red the norm κ_2 (see theorem 11). Vertices of the κ_2 unit ball constitute the $\widetilde{\mathcal{A}}_{2,1}$ set (see definition 8). The set $\widetilde{\mathcal{A}}_{2,1}$ belongs to the unit spheres of all three norms (see proposition 17).

We can deduce from it that the (k,q)-trace norm and (k,q)-CUT are nuclear norms, associated to particular vector norms:

Theorem 11 *1. The (k,q)-trace norm is the nuclear norm induced by θ_k on $\mathbb{R}^{m_1}$ and θ_q on $\mathbb{R}^{m_2}$, where for any $j \geq 1$, θ_j is the j-support norm introduced by Argyriou et al. (2012).*

2. The (k,q)-CUT norm is the nuclear norm induced by κ_k on $\mathbb{R}^{m_1}$ and κ_q on $\mathbb{R}^{m_2}$, where for any $j \geq 1$:

$$\kappa_j(w) = \frac{1}{\sqrt{j}} \max\left(\|w\|_\infty, \frac{1}{j}\|w\|_1 \right) . \tag{13}$$

For the sake of completeness, let us recall the closed-form expression of the k-support norm θ_k shown by Argyriou et al. (2012). For any vector $w \in \mathbb{R}^p$, let $\bar{w} \in \mathbb{R}^p$ be the vector obtained by sorting the entries of w by decreasing order of absolute values. Then it holds that

$$\theta_k(w) = \left\{ \sum_{i=1}^{k-r-1} |\bar{w}_i|^2 + \frac{1}{r+1} \left(\sum_{i=k-r}^{p} |\bar{w}_i| \right)^2 \right\}^{\frac{1}{2}} , \tag{14}$$

where $r \in \{0, \cdots, k-1\}$ is the unique integer such that $|\bar{w}_{k-r-1}| > \frac{1}{r+1}\sum_{i=k-r}^{p} |\bar{w}_i| \geq |\bar{w}_{k-r}|$, and where by convention $|\bar{w}_0| = \infty$.
Of course, Theorem 11 implies that in the vector case ($m_2 = 1$), the (k,q)-trace norm is simply equal to θ_k and the (k,q)-CUT norm is equal to κ_k. A representation of the "sharp edges" of unit balls of θ_k, κ_k and a appropriately scaled ℓ_1 norm can be found in Figure 1 for the case $m_1 = 3$ and $k = 2$. In addition, the following results shows that the dual norms of θ_k and κ_k have simple explicit forms:

Proposition 12 *The dual norms of θ_k and κ_k satisfy respectively:*

$$\theta_k^*(s) = \max_{I:|I|=k} \|s_I\|_2 \qquad \text{and} \qquad \kappa_k^*(s) = \frac{1}{\sqrt{k}} \max_{I:|I|=k} \|s_I\|_1 \,.$$

To conclude this section, let us observe that nuclear norms provide a natural framework to construct matrix norms from vector norms, and that other choices beyond θ_k and κ_k may lead to interesting norms for sparse matrix factorization. It is however known since Jameson (1987) (see also Bach, 2013; Bach et al., 2012) that the nuclear norm induced by vector ℓ_1-norm is simply the ℓ_1 of the matrix which fails to induce low rank (except in the very sparse case). However Bach et al. (2012) proposed nuclear norms associated with vectors norms that are similar to the elastic net penalty.

3 Learning matrices with sparse factors

In this section, we briefly discuss how the (k,q)-trace norm and (k,q)-CUT norm can be used to attack various problems involving estimation of sparse low-rank matrices.

3.1 Denoising

Suppose $X \in \mathbb{R}^{m_1 \times m_2}$ is a noisy observation of a low-rank matrix with sparse factors, assumed to have low (k,q)-rank. A natural convex formulation to recover the noiseless matrix is to solve:

$$\min_Z \frac{1}{2} \|Z - X\|_{\text{Fro}}^2 + \lambda \Omega_{k,q}(Z) \,, \tag{15}$$

where λ is a parameter to be tuned. Note that in the limit when $\lambda \to 0$, one simply obtains a soft-(k,q)-SVD of X.

3.2 Bilinear regression

More generally, given some empirical risk $\mathcal{L}(Z)$, it is natural to consider formulations of the form

$$\min_Z \mathcal{L}(Z) + \lambda \Omega_{k,q}(Z)$$

to learn matrices that are a priori assumed to have a low (k,q)-rank. A particular example is bilinear regression, where, given two inputs $x \in \mathbb{R}^{m_1}$ and $x' \in \mathbb{R}^{m_2}$, one observes as output a noisy version of $y = x^\top Z x'$. Assuming that Z has low (k,q)-rank means that the noiseless response is a sum of a small number of terms, each involving only a small number of features from either of the input vectors. To estimate such a model from observations $(x_i, x_i', y_i)_{i=1,\dots,n}$, one can consider the following convex formulation:

$$\min_Z \sum_{i=1}^n \ell\left(x_i^\top Z x_i', y_i\right) + \lambda \Omega_{k,q}(Z) \,, \tag{16}$$

where ℓ is a loss function. A particular instance of (16) of interest is the quadratic regression problem, where $m_1 = m_2$ and $x_i = x_i'$ for $i = 1, \dots, n$. Quadratic regression combined with additional constraints on Z is closely related to phase retrieval (Candès et al., 2013). It should be noted that if ℓ is the least-square loss, (16) can be rewritten in the form

$$\min_Z \frac{1}{2} \|\mathcal{X}(Z) - y\|_2^2 + \lambda \Omega_{k,q}(Z) \,,$$

where $\mathcal{X}(Z)$ is a linear transformation of Z, so that the problem is from the point of view of the parameter Z a linear regression with a well chosen feature map.

3.3 Subspace clustering.

In subspace clustering, one assumes that the data can be clustered in such a way that the points in each cluster belong to a low dimensional space. If we have a design matrix $X \in \mathbb{R}^{n\times p}$ with each row corresponding to an observation, then the previous assumption means that if $X^{(j)} \in \mathbb{R}^{n_j \times p}$ is a matrix formed by the rows of cluster j, there exist a low rank matrix $Z^{(j)} \in \mathbb{R}^{n_j \times n_j}$ such that $Z^{(j)}X^{(j)} = X^{(j)}$. This means that there exists a block-diagonal matrix Z such that $ZX = X$ with low-rank diagonal blocks. This idea, exploited recently by Wang et al. (2013) implies that Z is a sum of low rank sparse matrices; and this property still holds if the clustering is unknown. We therefore suggest that if all subspaces are of dimension k, Z may be estimated via

$$\min_Z \Omega_{k,k}(Z) \quad \text{s.t.} \quad ZX = X \ .$$

3.4 Sparse PCA

In sparse PCA (d'Aspremont et al., 2007; Witten et al., 2009; Zou et al., 2006), one tries to approximate an empirical covariance matrix $\hat{\Sigma}_n$ by a low-rank matrix with sparse factors. Although this is similar to the denoising problem discussed in Section 3.1, one may wish in addition that the estimated sparse low-rank matrix be symmetric and positive semi-definite (PSD), in order to represent a plausible covariance matrix. This suggests to formulate sparse PCA as follows:

$$\min_Z \left\{ \left\| \hat{\Sigma}_n - Z \right\|_{\text{Fro}} \ : \ (k,k)\text{-rank}(Z) \leq r \text{ and } Z \succeq 0 \right\} , \tag{17}$$

where k is the maximum number of non-zero coefficient allowed in each principal direction. In contrast to sequential approaches that estimate the principal components one by one (Mackey, 2009), this formulation requires to find simultaneously a set of factors which are complementary to one another in order to explain as much variance as possible. A natural convex relaxation of (17) is

$$\min_Z \left\{ \frac{1}{2} \left\| \hat{\Sigma}_n - Z \right\|^2_{\text{Fro}} + \lambda \Omega_{k,k}(Z) \ : \ Z \succeq 0 \right\} , \tag{18}$$

where λ is a parameter that controls in particular the rank of the approximation.
However, although the solution of (18) is always PSD, its soft-(k,k)-SVD $Z = \sum_{i=1}^r c_i a_i b_i^\top$ may not be composed of symmetric matrices (if $a_i \neq b_i$), and even if $a_i = b_j$ the corresponding c_i may be negative, as the following proposition shows:

Proposition 13 *1. The (k,k)-SVD of a PSD matrix is not necessarily a sum of symmetric terms.*

2. Some PSD matrices cannot be written as a positive combination of rank one (k,k)-sparse matrices, even for $k > 1$.

This may be unappealing, as one would like to interpret the successive rank-1 matrices as covariance matrices over a subspace that explain some of the total variance. One may therefore prefer a decomposition with less sparse or more factors, potentially capturing less variance.
One solution is to replace $\Omega_{k,k}$ in (18) by another penalty which directly imposes symmetric factors with non-negative weights. This is easily obtained by replacing the set of atoms $\mathcal{A}_{k,k}$ in Definition 4 by $\mathcal{A}_{k,\succeq} = \{aa^\top, a \in \mathcal{A}_k\}$, and considering the corresponding atomic norm which we denote by $\Omega_{k,\succeq}$. To be precise, $\Omega_{k,\succeq}$ is not a norm but only a gauge because the set $\mathcal{A}_{k,\succeq}$ is not centrally symmetric. Instead of (18), it possible to use the following convex formulation of sparse PCA:

$$\min_Z \frac{1}{2} \left\| \hat{\Sigma}_n - Z \right\|^2_{\text{Fro}} + \lambda \Omega_{k,\succeq}(Z) \, . \tag{19}$$

By construction, the solution of (19) is not only PSD, but can be expanded as a sum of matrices $Z = \sum_{i=1}^{r} c_i a_i a_i^\top$, where for all $i = 1, \ldots, r$, the factor a_i is k-sparse and the coefficient c_i is positive. This formulation is therefore particularly relevant if $\hat{\Sigma}_n$ is believed to be a noisy matrix of this form. It should be noted however that, by Proposition 13, $\Omega_{k,\succeq}$ is infinite for some PSD matrices[2], which implies that some PSD matrices cannot be approximated well with this formulation.

3.5 NP-hard convex problems

Although the (k,q)-trace norm and related norms allow us to formulate several problems of sparse low-rank matrix estimation as convex optimization problems, it should be pointed out that this does not guarantee the existence of efficient computational procedures to solve them. Here we illustrate this with the special case of the best (k,q)-sparse and rank 1 approximation to a matrix, which turns out to be a NP-hard problem. Indeed, let us consider the three following optimization problems, which are equivalent since they return the same rank one subspace spanned by $ab^\top$:

$$\min_{(a,b,c)\in\mathcal{A}_k\times\mathcal{A}_q\times\mathbb{R}^+} \|X - cab^\top\|_{\mathrm{Fro}}^2 \ ; \qquad \max_{(a,b)\in\mathcal{A}_k\times\mathcal{A}_q} a^\top X b \ ; \qquad \max_{Z:\,\Omega_{k,q}(Z)\leq 1} \mathrm{tr}(XZ^\top)\,. \tag{20}$$

In particular, if $k = q$ and $X = \widehat{\Sigma}_n$ is an empirical covariance matrix, then the symmetric solutions of the problem considered are the solution to the following rank 1 SPCA problem

$$\max_z \left\{ z^\top \hat{\Sigma}_n z \ : \ \|z\|_2 = 1 \ , \ \|z\|_0 \leq k \right\}, \tag{21}$$

which it is known to be NP-hard (Moghaddam et al., 2008). This shows that, in spite of being a convex formulation involving the (k,q)-trace norm, the third formulation in (20) is actually NP-hard. In practice, we will propose heuristics in Section 6 to approximate the solution of convex optimization problems involving the (k,q)-trace norm.

4 Statistical properties of the (k,q)-trace norm and the (k,q)-CUT norm

In this section we study theoretically the benefits of using the new penalties $\Omega_{k,q}$ and $\widetilde{\Omega}_{k,q}$ to infer low-rank matrices with sparse factors, as suggested in Section 3, postponing the discussion of how to do it in practice to Section 5. Building upon techniques proposed recently to analyze the statistical properties of sparsity-inducing penalties, such as the ℓ_1 penalty or more general atomic norms, we investigate two approaches to derive statistical guarantees. In Section 4.1 we study the expected dual norm of some noise process, from which we can deduce upper bounds on the learning rate for least squares regression and a simple denoising task. In Section 4.2 we estimate the statistical dimension of objects of interest both in the matrix and vector cases and compare the asymptotic rates, which shed light on the power of the norms we study when used as convex penalties. The results in Section 4.1 are technically easier to derive and contain bounds for a matrix of arbitrary (k,q)-rank. The results provided in Section 4.2 rely on a more involved set of tools, they provide more powerful bounds but we do not derive results for matrices of arbitrary (k,q)-rank.

4.1 Performance of the (k,q)-trace norm in denoising

In this Section we consider the simple denoising setting (Section 3.1) where we wish to recover a low-rank matrix with sparse factors $Z^\star \in \mathbb{R}^{m_1\times m_2}$ from a noisy observation $Y \in \mathbb{R}^{m_1\times m_2}$ corrupted

[2]This is possible because $\Omega_{k,\succeq}$ is only a gauge and not a norm.

by additive Gaussian noise:

$$Y = Z^\star + \sigma G\,,$$

where $\sigma > 0$ and G is a random matrix with entries i.i.d. from $\mathcal{N}(0,1)$. Given a convex penalty $\Omega : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}$, we consider, for any $\lambda > 0$, the estimator

$$\hat{Z}^\lambda_\Omega \in \arg\min_Z \frac{1}{2}\|Z - Y\|^2_{\mathrm{Fro}} + \lambda\Omega(Z)\,.$$

The following result, valid for any norm Ω, provides a general control of the estimation error in this setting, involving the dual norm of the noise:

Lemma 14 *If $\lambda \geq \sigma\Omega^*(G)$ then*

$$\left\|\hat{Z}^\lambda_\Omega - Z^\star\right\|^2_{\mathrm{Fro}} \leq 4\lambda\Omega(Z^\star)\,.$$

This suggests to study the dual norm of a random noise matrix $\Omega^*(G)$ in order to derive a upper bound on the estimation error. The following result provides such upper bounds, in expectation, for the (k,q)-trace norm as well as the standard ℓ_1 and trace norms:

Proposition 15 *Let $G \in \mathbb{R}^{m_1 \times m_2}$ be a random matrix with entries i.i.d. from $\mathcal{N}(0,1)$. The expected dual norm of G for the (k,q)-trace norm, the ℓ_1 norm and the trace norm is respectively bounded by:*

$$\begin{aligned}
\mathbb{E}\,\Omega^*_{k,q}(G) &\leq 4\left(\sqrt{k\log\frac{m_1}{k} + 2k} + \sqrt{q\log\frac{m_2}{q} + 2q}\right)\,,\\
\mathbb{E}\,\|G\|^*_1 &\leq \sqrt{2\log(m_1 m_2)}\,,\\
\mathbb{E}\,\|G\|^*_* &\leq \sqrt{m_1} + \sqrt{m_2}\,.
\end{aligned} \tag{22}$$

To derive an upper bound in estimation errors from these inequalities, we consider for simplicity[3] the oracle estimate $\hat{Z}^{\mathrm{Oracle}}_\Omega$ equal to $\hat{Z}^\lambda_\Omega$ where $\lambda = \sigma\Omega^*(G)$. From Lemma 14 we immediately get the following control of the mean estimation error of the oracle estimator, for any penalty Ω:

$$\mathbb{E}\left\|\hat{Z}^{\mathrm{Oracle}}_\Omega - Z^\star\right\|^2_{\mathrm{Fro}} \leq 4\sigma\Omega(Z^\star)\mathbb{E}\,\Omega^*(G)\,. \tag{23}$$

We can now derive upper bounds in estimation errors for the different penalties in the so-called single spike model, where the signal $Z^\star$ consists of an atom $ab^\top \in \mathcal{A}_{k,q}$, and we observed a noisy matrix $Y = ab^\top + \sigma G$. Since for an atom $ab^\top \in \mathcal{A}_{k,q}$ while $\|ab^\top\|_1 \leq kq/\sqrt{(kq)} = \sqrt{kq}$, $\Omega_{k,q}(ab^\top) = \|ab^\top\|_* = 1$, we immediately get the following by plugging the upper bounds of Proposition 15 into (23):

Corollary 16 *When $Z^\star \in \mathcal{A}_{k,q}$ is an atom, the expected errors of the oracle estimators using respectively the (k,q)-trace norm, the ℓ_1 norm and the trace norm are respectively upper bounded by:*

$$\begin{aligned}
\mathbb{E}\left\|\hat{Z}^{Oracle}_{\Omega_{k,q}} - Z^\star\right\|^2_{\mathrm{Fro}} &\leq 8\;\sigma\;\left(\sqrt{k\log\frac{m_1}{k} + 2k} + \sqrt{q\log\frac{m_2}{q} + 2q}\right)\,,\\
\mathbb{E}\left\|\hat{Z}^{Oracle}_{1} - Z^\star\right\|^2_{\mathrm{Fro}} &\leq 2\sigma\|Z^\star\|_1\sqrt{2\log(m_1 m_2)} \leq 2\sigma\sqrt{2kq\log(m_1 m_2)}\,,\\
\mathbb{E}\left\|\hat{Z}^{Oracle}_{*} - Z^\star\right\|^2_{\mathrm{Fro}} &\leq 2\sigma(\sqrt{m_1} + \sqrt{m_2})\,.
\end{aligned} \tag{24}$$

To make the comparison easy, orders of magnitudes of these upper bounds are gathered in Table 1 for the case where $Z^\star \in \widetilde{\mathcal{A}}_{k,q}$, and for the case where $m_1 = m_2 = m$ and $k = q = \sqrt{m}$. In the later

[3] Similar bounds could be derived with large probability for the non-oracle estimator by controlling the deviations of $\Omega^*(G)$ from its expectation.

Matrix norm	(k,q)-trace	trace	ℓ_1
$\Omega(Z^\star)\mathbb{E}\ \Omega^*(G)$	$\sqrt{k\log\frac{m_1}{k}}+\sqrt{q\log\frac{m_2}{q}}$	$\sqrt{m_1}+\sqrt{m_2}$	$\sqrt{kq\log(m_1m_2)}$
$k=\sqrt{m}$	$m^{1/4}\sqrt{\log m}$	$\sqrt{m}$	$\sqrt{m\log m}$

Table 1: Various norms mean square error in denoising an atom $ab^\top \in \widetilde{\mathcal{A}}_{k,q}$ corrupted with unit variance Gaussian noise. The column "$k=\sqrt{m}$" corresponds to the order of magnitudes in the regime where $m=m_1=m_2$ and $k=q=\sqrt{m}$.

case, we see in particular that the (k,q)-trace norm has a better rate than the ℓ_1 and trace norms, in $m^{\frac{1}{4}}$ instead of $m^{\frac{1}{2}}$ (up to logarithmic terms). Note that the largest value of $\|Z^\star\|_1$ is reached when $Z^\star \in \widetilde{\mathcal{A}}_{k,q}$ and equals $\sqrt{kq}$. By contrast, when $Z^\star \in \mathcal{A}_{k,q}$ gets far from $\widetilde{\mathcal{A}}_{k,q}$ elements then the expected error norm diminishes for the ℓ_1-penalized denoiser $\hat{Z}_1^{\text{Oracle}}$ reaching $\sigma\sqrt{2\ \log(m_1m_2)}$ on $e_1e_1^\top$ while not changing for the two other norms.

Obviously the comparison of upper bounds is not enough to conclude to the superiority of (k,q)-trace norm and, admittedly, the problem of denoising considered here is a special instance of linear regression in which the design matrix is the identity, and, since this is a case in which the design is trivially incoherent, it is possible to obtain fast rates for decomposable norms such as the ℓ_1 or trace norm (Negahban et al., 2012); however, slow rates are still valid in the presence of an incoherent design, or when the signal to recover is only weakly sparse, which is not the case for the fast rates. Moreover, the result proved here is valid for matrices of rank greater than 1. We present in the next section more involved results, based on lower and upper bounds on the so-called statistical dimension of the different norms (Amelunxen et al., 2013), a measure which is closely related to Gaussian widths.

4.2 Performance through the statistical dimension

Powerful results from asymptotic geometry have recently been used by Amelunxen et al. (2013); Chandrasekaran et al. (2012); Foygel and Mackey (2014); Oymak et al. (2013) to quantify the statistical power of a convex nonsmooth regularizer used as a constraint or penalty. These results rely essentially on the fact that if the tangent cone[4] of the regularizer at a point of interest Z is thiner, then the regularizer is more efficient at solving problems of denoising, demixing and compressed sensing of Z. The gain in efficiency can be quantified by appropriate measures of width of the tangent cone such as the Gaussian width of its intersection with a unit Euclidean ball (Chandrasekaran et al., 2012), or the closely related concept of *statistical dimension* of the cone, proposed by Amelunxen et al. (2013). In this section, we study the statistical dimensions induced by different matrix norms in order to compare their theoretical properties for exact or approximate recovery of sparse low-rank matrices. In particular, we will consider the norms $\Omega_{k,q}$, $\widetilde{\Omega}_{k,q}$ and linear combinations of the ℓ_1 and trace norms, which have been used in the literature to infer sparse low-rank matrices (Oymak et al., 2012; Richard et al., 2012). For convenience we therefore introduce the notation Γ_μ for the norm that linearly interpolates between the trace norm and the (scaled) ℓ_1 norm:

$$\forall \mu \in [0,1],\ \forall Z \in \mathbb{R}^{m_1\times m_2},\quad \Gamma_\mu(Z) := \frac{\mu}{\sqrt{kq}}\|Z\|_1 + (1-\mu)\|Z\|_*\ , \tag{25}$$

so that Γ_0 is the trace norm and Γ_1 is the ℓ_1 norm up to a constant[5].

[4] As detailed later, the tangent cone is the closure of the cone of descent directions.

[5] Note that the scaling ensures that $\Gamma_\mu(A)=1$ for $\mu\in[0,1]$ and $A\in\widetilde{\mathcal{A}}_{k,q}$.

4.2.1 The statistical dimension and its properties

Let us first briefly recall what the statistical dimension of a convex regularizer $\Omega : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}$ refers to, and how it is related to efficiency of the regularizer to recover a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$. For that purpose, we first define the tangent cone $T_\Omega(Z)$ of Ω at Z as the closure of the cone of descent directions, *i.e.*,

$$T_\Omega(Z) := \overline{\bigcup_{\tau > 0} \{H \in \mathbb{R}^{m_1 \times m_2} \ : \ \Omega(Z + \tau H) \leq \Omega(Z)\}}. \tag{26}$$

The statistical dimension $\mathfrak{S}(Z, \Omega)$ of Ω at Z can then be formally defined as

$$\mathfrak{S}(Z, \Omega) := \mathbb{E}\left[\left\|\Pi_{T_\Omega(Z)}(G)\right\|_{\text{Fro}}^2\right], \tag{27}$$

where G is a random matrix with i.i.d. standard normal entries and $\Pi_{T_\Omega(Z)}(G)$ is the orthogonal projection of G onto the cone $T_\Omega(Z)$. The statistical dimension is a powerful tool to quantify the statistical performance of a regularizer in various contexts, as the following non-exhaustive list of results shows.

- **Exact recovery with random measurements.** Suppose we observe $y = \mathcal{X}(Z^\star)$ where $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ is a random linear map represented by random design matrices X_i $i = 1, \ldots, n$ having iid entries drawn from $\mathcal{N}(0, 1/n)$. Then Chandrasekaran et al. (2012, Corollary 3.3) shows that

$$\hat{Z} = \arg\min_Z \Omega(Z) \quad \text{s.th.} \quad \mathcal{X}(Z) = y \tag{28}$$

  is equal to $Z^\star$ with overwhelming probability as soon as $n \geq \mathfrak{S}(Z^\star, \Omega)$. In addition Amelunxen et al. (2013, Theorem II) show that a phase transition occurs at $n = \mathfrak{S}(Z^\star, \Omega)$ between a situation where recovery fails with large probability (for $n \leq \mathfrak{S}(Z^\star, \Omega) - \gamma\sqrt{m_1 m_2}$, for some $\gamma > 0$) to a situation where recovery works with large probability (for $n \geq \mathfrak{S}(Z^\star, \Omega) + \gamma\sqrt{m_1 m_2}$).

- **Robust recovery with random measurements.** Suppose we observe $y = \mathcal{X}(Z^\star) + \epsilon$ where $\mathcal{X}$ is again a random linear map, and in addition the observation is corrupted by a random noise $\epsilon \in \mathbb{R}^n$. If the noise is bounded as $\|\epsilon\|_2 \leq \delta$, then Chandrasekaran et al. (2012, Corollary 3.3) show that

$$\hat{Z} = \arg\min_Z \Omega(Z) \quad \text{s.th.} \quad \|\mathcal{X}(Z) - y\|_2 \leq \delta \tag{29}$$

  satisfies $\left\|\hat{Z} - Z^\star\right\|_{\text{Fro}} \leq 2\delta/\eta$ with overwhelming probability as soon as $n \geq (\mathfrak{S}(Z^\star, \Omega) + \frac{3}{2})/(1 - \eta)^2$.

- **Denoising.** Assume a collection of noisy observations $X_i = Z^\star + \sigma\epsilon_i$ for $i = 1, \cdots, n$ is available where $\epsilon_i \in \mathbb{R}^{m_1 \times m_2}$ has i.i.d. $\mathcal{N}(0, 1)$ entries, and let $Y = \frac{1}{n}\sum_{i=1}^n X_i$ denote their average. Chandrasekaran and Jordan (2013, Proposition 4) prove that

$$\hat{Z} = \arg\min_Z \|Z - Y\|_{\text{Fro}} \quad \text{s.th.} \quad \Omega(Z) \leq \Omega(Z^\star) \tag{30}$$

  satisfies $\mathbb{E}\left\|\hat{Z} - Z^\star\right\|_{\text{Fro}}^2 \leq \frac{\sigma^2}{n}\mathfrak{S}(Z^\star, \Omega)$.

- **Demixing.** Given two matrices $Z^\star, V^\star \in \mathbb{R}^{m_1 \times m_2}$, suppose we observe $y = \mathcal{U}(Z^\star) + V^\star$ where $\mathcal{U} : \mathbb{R}^{m_1 \times m_2} \mapsto \mathbb{R}^{m_1 \times m_2}$ is a random orthogonal operator. Given two convex functions $\Gamma, \Omega : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}$, Amelunxen et al. (2013, Theorem III) show that

$$(\hat{Z}, \hat{V}) = \underset{(Z,V)}{\arg\min}\, \Omega(Z) \quad \text{s.th.} \quad \Gamma(V) \leq \Gamma(V^\star) \quad \text{and} \quad y = \mathcal{U}(Z) + V$$

is equal to $(Z^\star, V^\star)$ with probability at least $1 - \eta$ provided that

$$\mathfrak{S}(Z^\star, \Omega) + \mathfrak{S}(V^\star, \Gamma) \leq m_1 m_2 - 4\sqrt{m_1 m_2 \log \frac{4}{\eta}}.$$

Conversely if $\mathfrak{S}(Z^\star, \Omega) + \mathfrak{S}(V^\star, \Gamma) \geq m_1 m_2 + 4\sqrt{m_1 m_2 \log \frac{4}{\eta}}$, the demixing fails with probability at least $1 - \eta$.

4.2.2 Some cone inclusions and their consequences

In this and subsequent sections, we wish to compare the behavior of $\Omega_{k,q}$ and $\widetilde{\Omega}_{k,q}$ and Γ_μ, as defined in (25). Before estimating and comparing the statistical dimensions of these norms, which requires rather technical proofs, let us first show through simple geometric arguments that for a number of matrices, the tangent cones of the different norms are actually nested. This will allow us to derive deterministic improvement in performance when a norm is used as regularizer instead of another, which should be contrasted with the kind of guarantees that will be derived from bounds on the statistical dimension and which are typically statements holding with very high probability. The results in this section are proved in Appendix C.

Proposition 17 *The norms considered satisfy the following equalities and inequalities:*

$$\forall \mu \in [0,1],\ \forall Z \in \mathbb{R}^{m_1 \times m_2}, \qquad \Gamma_\mu(Z) \leq \Omega_{k,q}(Z) \leq \widetilde{\Omega}_{k,q}(Z),$$
$$\forall \mu \in [0,1],\ \forall A \in \widetilde{\mathcal{A}}_{k,q}, \qquad \Gamma_\mu(A) = \Omega_{k,q}(A) = \widetilde{\Omega}_{k,q}(A) = 1.$$

Put informally, the unit balls of $\widetilde{\Omega}_{k,q}$, $\Omega_{k,q}$ and of all convex combinations of the trace norm and the scaled ℓ_1-norm are nested and meet for matrices in $\widetilde{\mathcal{A}}_{k,q}$. This property is illustrated in the vector case (for $\mu = 1$) on Figure 1. In fact $\widetilde{\mathcal{A}}_{k,q}$ is a subset of the extreme points of the unit norms of all those norms except for the scaled ℓ_1-norm (corresponding to the case $\mu = 1$). Given that the unit balls meet on $\widetilde{\mathcal{A}}_{k,q}$ and are nested, their tangent cones on $\widetilde{\mathcal{A}}_{k,q}$ must also be nested:

Corollary 18 *The following nested inclusions of tangent cones hold:*

$$\forall \mu \in [0,1],\ \forall A \in \widetilde{\mathcal{A}}_{k,q}, \qquad T_{\Gamma_\mu}(A) \supset T_{\Omega_{k,q}}(A) \supset T_{\widetilde{\Omega}_{k,q}}(A)\,. \tag{31}$$

As a consequence, for any $A \in \widetilde{\mathcal{A}}_{k,q}$, the statistical dimensions of the different norms satisfy:

$$\mathfrak{S}(A, \widetilde{\Omega}_{k,q}) \leq \mathfrak{S}(A, \Omega_{k,q}) \leq \mathfrak{S}(A, \Gamma_\mu)\,. \tag{32}$$

As reviewed in Section 4.2.1, statistical dimensions provide estimates for the performance of the different norms in different contexts. Plugging (32) in these results shows that to estimate an atom in $\widetilde{\mathcal{A}}_{k,q}$, using $\widetilde{\Omega}_{k,q}$ is at least as good as using $\Omega_{k,q}$ which itself is at least as good as using any convex combination of the ℓ_1 and trace norms.
Note that the various statements in Section 4.2.1 provide upper bounds on the performance of the different norms, with are guarantees that are either probabilistic or hold in expectation. In fact, the inclusion of the tangent cones (31) and a fortiori the tangential inclusion of the unit balls imply much stronger results since it can also lead some deterministic statements, such as the following:

Corollary 19 (Improvement in exact recovery) *Consider the problem of exact recovery of a matrix $Z^* \in \widetilde{\mathcal{A}}_{k,q}$ from random measurements $y = \mathcal{X}(Z^*)$ by solving (28) with the different norms. For any realization of the random measurements, exact recovery with Γ_μ for any $0 \leq \mu \leq 1$ implies exact recovery with $\Omega_{k,q}$ which itself implies exact recovery with $\widetilde{\Omega}_{k,q}$.*

Note that in the vector case ($m_2 = 1$), where the (k,q)-trace norm $\Omega_{k,1}$ boils down to the k-support norm θ_k, the tangent cone inclusion (31) is not always strict:

Proposition 20 *For any $a \in \widetilde{\mathcal{A}}_k^m$, $T_{\Gamma_1}(a) = T_{\theta_k}(a)$.*

In words, the tangent cone of the ℓ_1 norm and of the the k-support norm are equal on k-sparse vectors with constant non-zero entries, which can be observed in Figure 1. This suggests that, in the vector case, the k-support norm is not better than the ℓ_1 norm to recover such constant sparse k-vectors.

4.2.3 Bounds on the statistical dimensions

The results presented in Section 4.2.2 apply only to a very specific set of matrices ($\widetilde{\mathcal{A}}_{k,q}$), and do not characterize quantitatively the relative performance of the different norms. In this Section, we turn to more explicit estimations of the statistical dimension of the different norms at atoms in $\widetilde{\mathcal{A}}_{k,q}$ and $\mathcal{A}_{k,q}$.

We consider first the statistical dimension of the (k,q)-CUT norm $\widetilde{\Omega}_{k,q}$ on its atoms $\widetilde{\mathcal{A}}_{k,q}$. The unit ball of $\widetilde{\Omega}_{k,q}$ is a vertex-transitive polytope with $2^{k+q}\binom{m_1}{k}\binom{m_2}{q}$ vertices. As a consequence, it follows immediately from Corollary 3.14 in Chandrasekaran et al. (2012) and from the upper bound $\log\binom{m}{k} \leq k(1+\log(m/k))$, that[6]

Proposition 21 *For any $A \in \widetilde{\mathcal{A}}_{k,q}$, we have*

$$\mathfrak{S}(A, \widetilde{\Omega}_{k,q}) \;\leq\; 16(k+q) + 9\left(k \log \frac{m_1}{k} + q \log \frac{m_2}{q}\right).$$

Upper bounding the statistical dimension of the (k,q)-trace norm on its atoms $\mathcal{A}_{k,q}$ requires more work. First, atoms with very small coefficients are likely to be more difficult to estimate than atoms with large coefficients only. In the vector case, for example, it is known that the recovery of a sparse vector β with support I_0 depends on its smallest coefficient $\beta_{\min} = \min_{i \in I_0} \beta_i^2$ (Wainwright, 2009). The ratio between $\beta_{\min}$ and the noise level can be thought of as the worst signal-to-noise ratio for the signal β. We generalize this idea to atoms in $\mathcal{A}_{k,q}$ as follows.

Definition 22 (Atom strength) *Let $A = ab^\top \in \mathcal{A}_{k,q}$ with $I_0 = supp(a)$ and $J_0 = supp(b)$. Denote $a_{\min}^2 = \min_{i \in I_0} a_i^2$ and $b_{\min}^2 = \min_{j \in J_0} b_j^2$. The atom strength $\gamma(a,b) \in (0,1]$ is*

$$\gamma(a,b) := (k\, a_{\min}^2) \wedge (q\, b_{\min}^2).$$

Note that the atoms with maximal strength value 1 are the elements of $\widetilde{\mathcal{A}}_{k,q}$. With this notion in hand we can now formulate an upper bound on the statistical dimension of $\Omega_{k,q}$:

Proposition 23 *For $A = ab^\top \in \mathcal{A}_{k,q}$ with strength $\gamma = \gamma(a,b)$, we have*

$$\mathfrak{S}(A, \Omega_{k,q}) \leq \frac{322}{\gamma^2}(k+q+1) + \frac{160}{\gamma}(k \vee q) \log(m_1 \vee m_2)\,. \tag{33}$$

[6] This result is actually stated informally for the special case of $k = q = \sqrt{m} =$ with $m = m_1 = m_2$ in the context of a discussion of the planted clique problem in Chandrasekaran and Jordan (2013).

Matrix norm	$\mathfrak{S}$	$k=\sqrt{m}$	Vector norm	$\mathfrak{S}$
(k,q)-trace	$\mathcal{O}((k\vee q)\log(m_1\vee m_2))$	$\mathcal{O}(\sqrt{m}\log m)$	k-support	$\Theta(k\log\frac{p}{k})$
(k,q)-cut	$\mathcal{O}(k\log\frac{m_1}{k}+q\log\frac{m_2}{q})$	$\mathcal{O}(\sqrt{m}\log m)$	κ_k	$\Theta(k\log\frac{p}{k})$
ℓ_1	$\Theta(kq\ \log\frac{m_1m_2}{kq})$	$\Theta(m\log m)$	ℓ_1	$\Theta(k\log\frac{p}{k})$
trace-norm	$\Theta(m_1+m_2)$	$\Theta(m)$	ℓ_2	p
ℓ_1+trace-n.	$\Omega\big(kq\wedge(m_1+m_2)\big)$	$\Theta(m)$	elastic net	$\Theta(k\log\frac{p}{k})$
"cut-norm"	$\mathcal{O}(m_1+m_2)$	$\mathcal{O}(m)$	ℓ_∞	p

Table 2: Order of magnitude of the statistical dimension of different matrix norms for elements of $\widetilde{\mathcal{A}}_{k,q}$ (left) and of their vector norms counterpart for elements of $\widetilde{\mathcal{A}}_k^p$ (right). The ℓ_1 norm here is the element-wise ℓ_1 norm. The column "$k=\sqrt{m}$" corresponds to the case of the planted clique problem where $m=m_1=m_2$ and $k=q=\sqrt{m}$. We use usual Landau notation with $f=\Theta(g)$ for $(f=\mathcal{O}(g))\&(g=\mathcal{O}(f))$ and $f=\Omega(g)$ for $g=\mathcal{O}(f)$. The absence of Landau notation means that the computation is exact.

Note that the upper bounds obtained on atoms of $\widetilde{\mathcal{A}}_{k,q}$ for $\widetilde{\Omega}_{k,q}$ (Proposition 21) and $\Omega_{k,q}$ (Proposition 23, with $\gamma=1$) have the same rate up to $k\log k+q\log q$ which is negligible compared to $k\log m_1+q\log m_2$ when $k\ll m_1$ and $q\ll m_2$. Note that once the support is specified, the number of degrees of freedom for elements of $\widetilde{\mathcal{A}}_{k,q}$ is $k+q-1$, which is matched up to logarithmic terms. It is interesting to compare these estimates to the statistical dimension of the ℓ_1 norm, the trace norm, and their combinations Γ_μ. Table 2 summarizes the main results. The statistical dimension the ℓ_1 norm on atoms in $\widetilde{\mathcal{A}}_{k,q}$ is of order $kq\log(m_1m_2/(kq))$, which is worse than the statistical dimensions of $\Omega_{k,q}$ and $\widetilde{\Omega}_{k,q}$ by a factor $k\wedge q$. On $\mathcal{A}_{k,q}$, though, the statistical dimension of $\Omega_{k,q}$ increases when the atom strength decreases, while the statistical dimension of the ℓ_1 norm is independent of it and even decreases when the size of the support decreases. As for the trace norm alone, its statistical dimension is at least of order m_1+m_2, which is unsurprisingly much worse that the statistical dimensions of $\Omega_{k,q}$ and $\widetilde{\Omega}_{k,q}$ since it does not exploit the sparsity of the atoms. Finally, regarding the combination Γ_μ of the ℓ_1 norm and of the trace norm, Oymak et al. (2012) has shown that it does not improve rates up to constants over the best of the two norms. More precisely, we can derive from Oymak et al. (2012, Theorem 3.2) the following result

Proposition 24 *There exists $M>0$ and $C>0$ such that for any $m_1,m_2,k,q\geq M$ with $m_1/k\geq M$ and $m_2/q\geq M$, for any $A\in\mathcal{A}_{k,q}$ and for any $\mu\in[0,1]$, the following holds:*

$$\mathfrak{S}\left(A,\Gamma_\mu\right)\geq C\,\zeta(a,b)\left(\left(kq\right)\wedge(m_1+m_2-1)\right)-2\,,$$

with

$$\zeta(a,b)=1-\Big(1-\frac{\|a\|_1^2}{k}\Big)\Big(1-\frac{\|b\|_1^2}{q}\Big)\,.$$

Note that $\zeta(a,b)\leq 1$ with equality if either $a\in\widetilde{\mathcal{A}}_k^{m_1}$ or $b\in\widetilde{\mathcal{A}}_q^{m_2}$, so in particular $\zeta(a,b)=1$ for $ab^\top\in\widetilde{\mathcal{A}}_{k,q}$. In that case, we see that, as stated by Oymak et al. (2012), Γ_μ does not bring any improvement over the ℓ_1 and trace norms taken imdividually, and in particular has a worse statistical dimension than $\Omega_{k,q}$ and $\widetilde{\Omega}_{k,q}$.

4.2.4 The vector case

We have seen in Section 4.2.3 that the statistical dimension of the (k,q)-trace norm and of the (k,q)-CUT norm were smaller than that of the ℓ_1 and the trace norms, and of their combinations,

meaning that theoretically they are more efficient regularizers to recover rank-one sparse matrices. In this section, we look more precisely at these properties in the vector case ($m_2 = q = 1$), and show that, surprisingly, the benefits are lost in this case.

Remember that, in the vector case, $\Omega_{k,q}$ boils down to the k-support norm θ_k (14), while $\widetilde{\Omega}_{k,q}$ boils down to the norm κ_k (13). For the later, we can upper bound the statistical dimension at a k-sparse vector by specializing Proposition 21 to the vector case, and also derive a specific lower bound as follows:

Proposition 25 *For any k-sparse vector $a \in \widetilde{\mathcal{A}}_k^p$,*

$$\frac{k}{2\pi} \log\left(\frac{p-k}{k+1}\right) \leq \mathfrak{S}(a, \kappa_k) \leq 9k \log \frac{p}{k} + 16(k+1)\,.$$

From the explicit formulation of θ_k (14) we can derive an upper bound of the statistical dimension of θ_k on any sparse vector with at least k non-zero coefficients:

Proposition 26 *For any $s \geq k$, the statistical dimension of the k-support norm θ_k at an s-sparse vector $w \in \mathbb{R}^p$ is bounded by*

$$\mathfrak{S}(w, \theta_k) \leq \frac{5}{4}s + 2\left\{\frac{(r+1)^2 \|\tilde{w}_{I_2}\|_2^2}{\|\tilde{w}_{I_1}\|_1^2} + |I_1|\right\} \log \frac{p}{s}\,, \tag{34}$$

where $\tilde{w} \in \mathbb{R}^p$ denotes the vector with the same entries as w sorted by decreasing absolute values, r is as defined in equation (14), $I_2 = [1, k-r-1]$ and $I_1 = [k-r, s]$. In particular, when $s = k$, the following holds for any atom $a \in \mathcal{A}_k^p$ with strength $\gamma = k a_{\min}^2$:

$$\mathfrak{S}(a, \theta_k) \leq \frac{5}{4}k + \frac{2k}{\gamma} \log \frac{p}{k}\,. \tag{35}$$

We note that (35) has the same rate but tighter constants than the general upper bound (33) specialized to the vector case. In particular, this suggests that the γ^{-2} term in (35) may not be required. In the lasso case ($k = 1$), we recover the standard bound (Chandrasekaran et al., 2012):

$$\mathfrak{S}(w, \theta_k) \leq \frac{5}{4}s + 2s \log \frac{p}{s}\,, \tag{36}$$

which is also reached by θ_k on an atom $a \in \widetilde{\mathcal{A}}_k^p$ because in that case $\gamma = 1$ in (35). On the other hand, for general atoms in $\mathcal{A}_k^p$ the upper bound (35) is always worse than the upper bound for the standard Lasso (36), and more generally the upper bound for general sparse vectors (34) is also never better than the one for the Lasso. Although these are only upper bounds, this raises questions on the utility of the k-support norm compared to the lasso to recover sparse vectors.

The statistical complexities of the different regularizers in the vector case are summarized in Table 2. We note that, contrary to the low-rank sparse matrix case, the ℓ_1-norm, the k-support norm, and the norm κ_k all have the same statistical dimension up to constants. Note that the tangent cone of the elastic net equals the tangent cone of the ℓ_1-norm in any point (because the tangent cone of the ℓ_2 norm is a half space that always contains the tangent cone of the ℓ_1-norm) so that the elastic net has always the exact same statistical dimension as the ℓ_1-norm.

5 Algorithms

As seen in Section 3, many problems involving sparse low-rank matrix estimation can be formulated as optimization problems of the form:

$$\min_{Z \in \mathbb{R}^{m_1 \times m_2}} \mathcal{L}(Z) + \lambda \Omega_{k,q}(Z). \tag{37}$$

Unfortunately, although convex, this problem may be computationally challenging (Section 3.5). In this section, we present a working set algorithm to approximately solve such problems in practice when $\mathcal{L}$ is differentiable.

5.1 A working set algorithm

Given a set $\mathcal{S} \subset \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}$ of pairs of row and column subsets, let us consider the optimization problem:

$$\min_{(Z^{(IJ)})_{(I,J)\in\mathcal{S}}} \left\{ \mathcal{L}\left(\sum_{(I,J)\in\mathcal{S}} Z^{(IJ)} \right) + \lambda \sum_{(I,J)\in\mathcal{S}} \left\| Z^{(IJ)} \right\|_* \ : \ \forall (I,J) \in \mathcal{S}, \operatorname{supp}(Z^{(IJ)}) \subset I \times J \right\}. \tag{$\mathcal{P}_\mathcal{S}$}$$

Let $(\widehat{Z^{(IJ)}})_{(I,J)\in\mathcal{S}}$ be a solution of this optimization problem. Then, by the characterization of $\Omega_{k,q}(Z)$ in (8), $Z = \sum_{(I,J)\in\mathcal{S}} \widehat{Z^{(IJ)}}$ is the solution of (37) when $\mathcal{S} = \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}$. Clearly, it is still the solution of (37) if $\mathcal{S}$ is reduced to the set of non-zero matrices $\widehat{Z^{(IJ)}}$ at optimality often called *active* components.

We propose to solve problem (37) using a so-called working set algorithm which solves a sequence of problems of the form ($\mathcal{P}_\mathcal{S}$) for a growing sequence of working sets $\mathcal{S}$, so as to keep a small number of non-zero matrices $Z^{(IJ)}$ throughout. Working set algorithms (Bach et al., 2011, Chap. 6) are typically useful to speed up algorithm for sparsity inducing regularizer; they have been used notably in the case of the overlapping group Lasso of Jacob et al. (2009) which is also naturally formulated *via* latent components.

To derive the algorithm we write the optimality condition for ($\mathcal{P}_\mathcal{S}$):

$$\forall (I,J) \in \mathcal{S}\,, \quad \nabla\mathcal{L}(Z)_{IJ} \in -\lambda \partial \left\| Z^{(IJ)} \right\|_* .$$

From the characterization of the subdifferential of the trace norm (Watson, 1992), writing $Z^{(IJ)} = U^{(IJ)}\Sigma^{(IJ)}V^{(IJ)}$ the SVD of $Z^{(IJ)}$, this is equivalent to, for all (I,J) in $\mathcal{S}$,

$$\text{either} \quad Z^{(IJ)} \neq 0 \quad \text{and} \quad \nabla\mathcal{L}(Z)_{IJ} = -\lambda \left(U^{(IJ)} {V^{(IJ)}}^\top + A \right)$$
$$\text{with } \|A\|_{\text{op}} \leq 1 \text{ and } AU^{(IJ)} = A^\top V^{(IJ)} = 0\,, \tag{38}$$
$$\text{or} \quad Z^{(IJ)} = 0 \quad \text{and} \quad \|\nabla\mathcal{L}(Z)]_{IJ}\|_{\text{op}} \leq \lambda\,. \tag{39}$$

The principle of the working set algorithm is to solve problem ($\mathcal{P}_\mathcal{S}$) for the current set $\mathcal{S}$ so that (38) and (39) are (approximately) satisfied for (I,J) in $\mathcal{S}$, and to check subsequently if there are any components not in $\mathcal{S}$ which violate (39). If not, this guarantees that we have found a solution to problem (37), otherwise the new pair (I,J) corresponding to the most violated constraint is added to $\mathcal{S}$ and problem ($\mathcal{P}_\mathcal{S}$) is initialized with the previous solution and solved again. The resulting algorithm is Algorithm 1 (where the routine SSVDTPI is described in the next section). Problem ($\mathcal{P}_\mathcal{S}$) is solved easily using the approximate block coordinate descent of Tseng and Yun (2009) (see

also Bach et al., 2011, Chap. 4), which consists in iterating proximal operators. The modifications to the algorithm to solve problems regularized by the norm $\Omega_{k,\succeq}$ are relatively minor (they amount to replace the trace norms by penalization of the trace of the matrices $Z^{(IJ)}$ and by positive definite cone constraints) and we therefore do not describe them here.
Determining efficiently which pair (I, J) possibly violates condition (39) is in contrast a more difficult problem that we discuss next.

Algorithm 1 Active set algorithm

Require: $\mathcal{L}$, tolerance $\epsilon > 0$, parameters λ, k, q
  Set $\mathcal{S} = \varnothing, Z = 0$
  while $c =$ `true` **do**
    Recompute optimal values of Z, $(Z^{(IJ)})_{(I,J)\in\mathcal{S}}$ for $(\mathcal{P}_\mathcal{S})$ using warm start
    $(I, J) \leftarrow$ `SSVDTPI`$(\nabla\mathcal{L}(Z), k, q, \epsilon)$
    if $\|[\nabla\mathcal{L}(Z)]_{I,J}\|_{\mathrm{op}} > \lambda$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(I, J)\}$
    else
      $c \leftarrow$ `false`
    end if
  end while
  return Z, $\mathcal{S}$, $(Z^{(IJ)})_{(I,J)\in\mathcal{S}}$

5.2 Finding new active components

Once $(\mathcal{P}_\mathcal{S})$ is solved for a given set $\mathcal{S}$, (38) and (39) are satisfied for all $(I, J) \in \mathcal{S}$. Note that (38) implies in particular that $\|\nabla\mathcal{L}(Z)]_{IJ}\|_{\mathrm{op}} = \lambda$ when $Z^{(IJ)} \neq 0$ at optimality. Therefore, (39) is also satisfied for all $(I, J) \notin \mathcal{S}$ if and only if

$$\max_{(I,J)\in\mathcal{G}_k^{m_1}\times\mathcal{G}_q^{m_2}} \|[\nabla\mathcal{L}(Z)]_{IJ}\|_{\mathrm{op}} \leq \lambda\,, \tag{40}$$

and if this is not the case then any (I, J) that violates this condition is a candidate to be included in $\mathcal{S}$. This corresponds to solving the following sparse singular value problem

$$\max_{a,b} \quad a^\top\nabla\mathcal{L}(Z)b \quad \text{s.t.} \quad ab^\top \in \mathcal{A}_{k,q}\,. \tag{k,q)-linRank-1}$$

This problem is unfortunately NP-hard since rank 1 sparse PCA problem is a particular instance of it (when $\nabla\mathcal{L}(Z)$ is replaced by a covariance matrix), and we therefore cannot hope to solve it exactly with efficient algorithms. Still, sparse PCA has been the object of a significant amount of research, and several relaxations and other heuristics have been proposed to solve it approximately. In our numerical experiments we use a truncated power iteration (TPI) method, also called TPower, GPower or CongradU in the PSD case (Journée et al., 2010; Luss and Teboulle, 2013; Yuan and Zhang, 2013), which has been proved recently by Yuan and Zhang (2013) to provide accurate solution in reasonable computational time under RIP type of conditions. Algorithm 2 provides a natural generalization of this algorithm to the non-PSD case. The algorithm follows the steps of a power method, the standard method for computing leading singular vectors of a matrix, with the difference that at each iteration a truncation step is use. We denote the truncation operator by T_k. It consists of keeping the k largest components (in absolute value) and setting the others to 0. Note that Algorithm 2 may fail to find a new active component for Algorithm 1 if it

Algorithm 2 SSVDTPI: Bi-truncated power iteration for (k,q)-linRank-1

Require: $A \in \mathbb{R}^{m_1 \times m_2}$, k, q and tolerance $\epsilon > 0$
  Pick a random initial point $b^{(0)} \sim \mathcal{N}(0, I_{m_2})$ and let
  while $|a^{(t)\top} A b^{(t)} - a^{(t-1)\top} A b^{(t-1)}| / |a^{(t-1)\top} A b^{(t-1)}| > \epsilon$ **do**
    $a \leftarrow Ab^{(t)}$ \\ `Power`
    $a \leftarrow T_k(a)$ \\ `Truncate`
    $b \leftarrow A^\top a$ \\ `Power`
    $b \leftarrow T_q(b)$ \\ `Truncate`
    $a^{(t+1)} \leftarrow a/\|a\|_2$ and $b^{(t+1)} \leftarrow b/\|b\|_2$ \\ `Normalize`
    $t \leftarrow t+1$
  end while
  $I \leftarrow \mathrm{Supp}(a^{(t)})$ and $J \leftarrow \mathrm{Supp}(b^{(t)})$
  return (I, J)

finds a local maximum of ((k,q)-linRank-1) smaller than λ, and therefore result in the termination of Algorithm 1 on a suboptimal solution. On the positive side, note that Algorithm 1 is robust to some errors of Algorithm 2. For instance, if an incorrect component is added to $\mathcal{S}$ at some iteration, but the correct components are identified later, Algorithm 1 will eventually shrink the incorrect components to 0. One of the causes of failure of TPI type of methods is the presence of a large local maximum in the sparse PCA problem corresponding to a suboptimal component; incorporating this component in $\mathcal{S}$ will reduce the size of that local maximum, thereby increasing the chance of selecting a correct component the next time around.

5.3 Computational cost

Note that when m_1, m_2 are large, solving $\mathcal{P}_\mathcal{S}$ involves the minimizations of trace norms of matrices of size $k \times q$ which, when k and q are small compared to m_1 and m_2 have low computational cost. The bottleneck for providing a computational complexity of the algorithm is the (k,q)-linRank-1 step. It has been proved by Yuan and Zhang (2013) that under some conditions the problem can be solved in linear time. If the conditions hold at every step of gradient, the overall cost of an iteration can be cast into the cost of evaluating the gradient and the evaluation of thin SVDs: $O(k^2 q)$. Evaluating the gradient has a cost dependent on the risk function $\mathcal{L}$. This cost for usual applications is $O(m_1 m_2)$. So assuming the RIP conditions required by Yuan and Zhang (2013) hold, the cost of Algorithm 2 is dominated by matrix-vector multiplications so of the order $O(m_1 m_2)$. The total cost of the algorithm for reaching a δ-accurate solution is therefore $O((m_1 m_2 + k^2 q)/\delta)$. However the worst case complexity of the algorithm is non-polynomial as (k,q)-linRank-1 is non-polynomial in general. We would like to point out that in our numerical experiments a warm start with singular vectors and multiple runs of the algorithm (k,q)-linRank-1 keeping track of the highest found variance has provided us a very fast and reliable solver. Further discussion on this step go beyond the scope of this work.

6 Numerical experiments

In this section we report experimental results to assess the performance of sparse low-rank matrix estimation using different techniques. We start in Section 6.1 with simulations aiming at validating the theoretical results on statistical dimension of $\Omega_{k,q}$ and assessing how they generalize to matrices

with (k,q)-rank larger than 1. In Section 6.2 we compare several techniques for sparse PCA on simulated data.

6.1 Empirical estimates of the statistical dimension.

In order to numerically estimate the statistical dimension $\mathfrak{S}(Z,\Omega)$ of a regularizer Ω at a matrix Z, we add to Z a random Gaussian noise matrix and observe $Y = Z + \sigma G$ where G has normal i.i.d. entries following $\mathcal{N}(0,1)$. We then denoise Y using (30) to form an estimate $\hat{Z}$ of Z. For small σ, the normalized mean-squared error (NMSE) defined as

$$\mathsf{NMSE}(\sigma) := \frac{\mathbb{E}\left\|\hat{Z} - Z\right\|_{\mathrm{Fro}}^2}{\sigma^2}$$

is a good estimate of the statistical dimension, since Oymak and Hassibi (2013) show that

$$\mathfrak{S}(Z,\Omega) = \lim_{\sigma \to 0} \mathsf{NMSE}(\sigma)\,.$$

Numerically, we therefore estimate $\mathfrak{S}(Z,\Omega)$ by taking $\sigma = 10^{-4}$ and measuring the empirical NMSE averaged over 20 repeats. We consider square matrices with $m_1 = m_2 = 1000$, and estimate the statistical dimension of $\Omega_{k,q}$, the ℓ_1 and the trace norms at different matrices Z. The constrained denoiser (30) has a simple close-form for the ℓ_1 and the trace norm. For $\Omega_{k,q}$, it can be obtained by a series of proximal projections (15) with different parameters λ until $\Omega_{k,q}(\hat{Z})$ has the correct value $\Omega_{k,q}(Z)$. Since the noise is small, we found that it was sufficient and faster to perform a soft-(k,q)-SVD of Y by solving (15) with a small λ, and then apply the ℓ_1 constrained denoiser to the set of soft-(k,q)-sparse singular values.

We first estimate the statistical dimensions of the three norms at an atom $Z \in \widetilde{\mathcal{A}}_{k,q}$, for different values of $k = q$. Figure 2 (top left) shows the results, which confirm the theoretical bounds summarized in Table 2. The statistical dimension of the trace norm does not depend on k, while that of the ℓ_1 norm increases almost quadratically with k and that of $\Omega_{k,q}$ increases linearly with k. As expected, $\Omega_{k,q}$ interpolates between the ℓ_1 norm (for $k = 1$) and the trace norm (for $k = m_1$), and outperforms both norms for intermediary values of k. This experiments therefore confirms that our upper bound (33) on $\mathfrak{S}(Z,\Omega_{k,q})$ captures the correct order in k, although the constants can certainly be much improved, and that Algorithm 1 manages, in this simple setting, to correctly approximate the solution of the convex minimization problem.

Second, we estimate the statistical dimension of $\Omega_{k,q}$ on matrices with (k,q)-rank larger than 1, a setting for which we proved no theoretical result. Figure 2 (top left) shows the numerical estimate of $\mathfrak{S}(Z,\Omega_{k,q})$ for matrices Z which are sums of r atoms in $\widetilde{\mathcal{A}}_{k,k}$ with non-overlapping support, for $k = 10$ and varying r. We observe that the increase in statistical dimension is roughly linear in the (k,q)-rank. For a fixed (k,q)-rank of 3, the bottom plots of Figure 2 compare the estimated statistical dimensions of the three regularizers on matrices Z which are sums of 3 atoms in $\widetilde{\mathcal{A}}_{k,k}$ with non-overlapping (bottom left) or overlapping (bottom right) supports. The shapes of the different curves are overall similar to the rank 1 case, although the performance of $\Omega_{k,q}$ degrades as the supports of atoms overlap. In both cases, $\Omega_{k,q}$ consistently outperforms the two other norms. Overall these experiments suggest that the statistical dimension of $\Omega_{k,q}$ at a linear combination of r atoms increases as $Cr\,(k \log m_1 + q \log m_2)$ where the coefficient C increases with the overlap among the supports of the atoms.

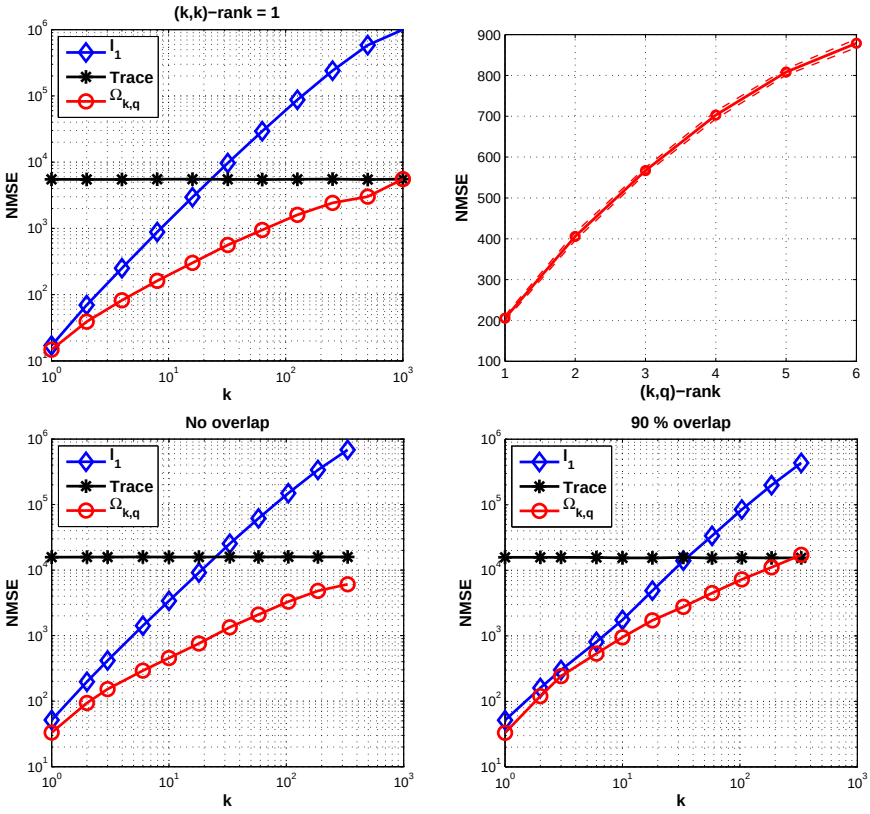

Figure 2: Estimates of the statistical dimensions of the ℓ_1, trace and $\Omega_{k,q}$ norms at a matrix $Z \in \mathbb{R}^{1000\times 1000}$ in different setting. Top left: Z is an atom in $\widetilde{\mathcal{A}}_{k,k}$ for different values of k. Top right: Z is a sum of r atoms in $\widetilde{\mathcal{A}}_{k,k}$ with non-overlapping support, with $k = 10$ and varying r. Bottom left: Z is a sum of 3 atoms in $\widetilde{\mathcal{A}}_{k,k}$ with non-overlapping support, for varying k. Bottom right: Z is a sum of 3 atoms in $\widetilde{\mathcal{A}}_{k,k}$ with overlapping support, for varying k.

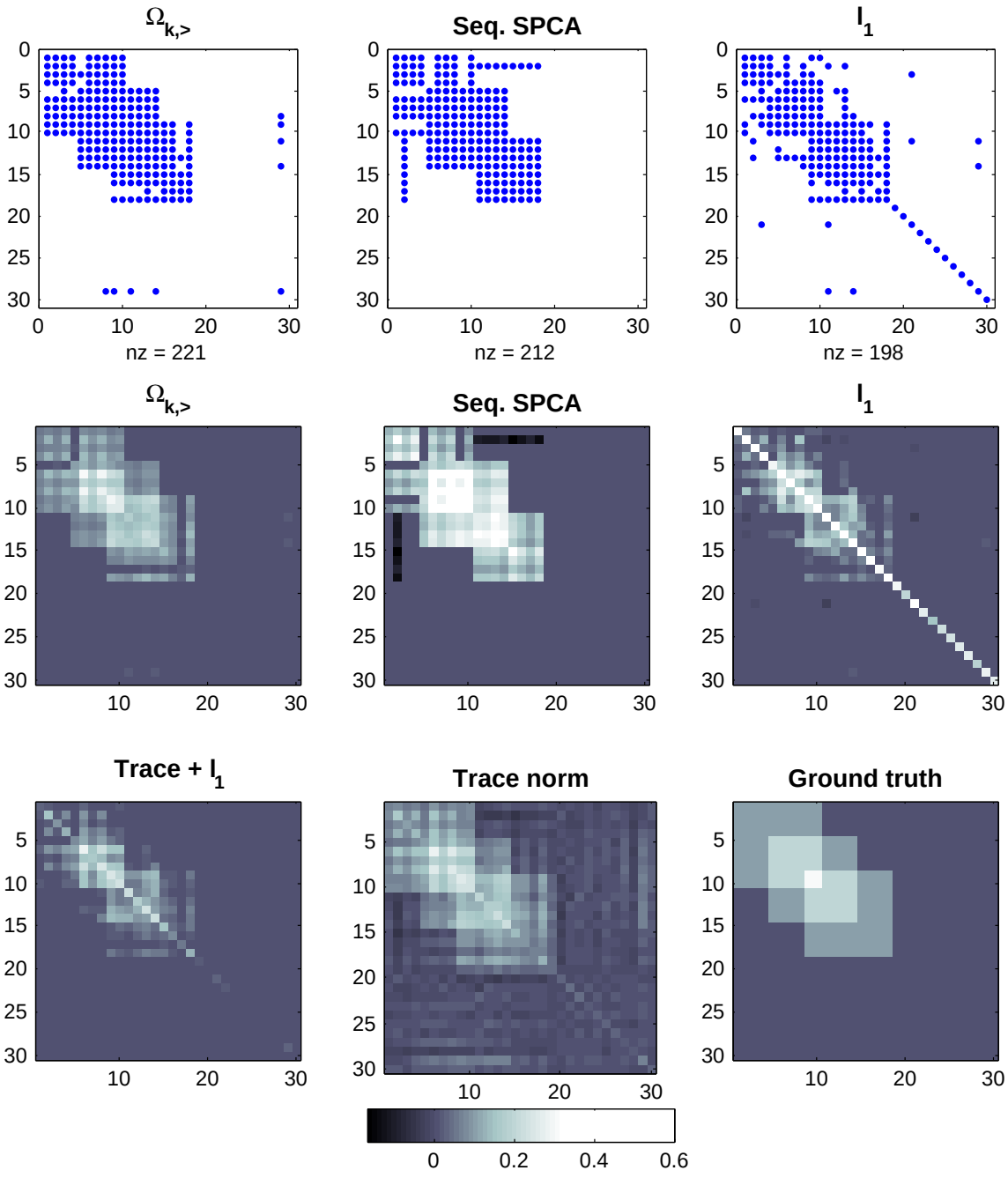

Figure 3: Sparse PCA example. The first row shows the supports found by our method (left) by sequential sparse PCA (middle) and element wise thresholding of the sample covariance matrix. Other plots contain heatmaps of the estimated covariance matrix using different methods, and the ground truth $\Sigma^\star$ in the lower right hand side.

6.2 Comparison of algorithms for sparse PCA

In this section we compare the performance of different algorithms in estimating a sparsely factored covariance matrix that we denote $\Sigma^\star$. The observed sample consists of n random vector vectors generated i.i.d. according to $\mathcal{N}(0, \Sigma^\star + \sigma^2 \mathrm{Id}_p)$, where (k,k)-rank$(\Sigma^\star) = 3$. The matrix $\Sigma^\star$ is formed by adding 3 blocks of rank 1, $\Sigma^\star = a_1 a_1^\top + a_2 a_2^\top + a_3 a_3^\top$, having all the same sparsity $\|a_i\|_0 = k = 10$, 3×3 overlaps and nonzero entries equal to $1/\sqrt{k}$. See Figure 3, bottom right plot for a representation of the ground truth $\Sigma^\star$. The noise level $\sigma = 0.8$ is set in order to make the signal to noise ratio below the level $\sigma = 1$ where a spectral gap appears and makes the spectral baseline (penalizing the trace of the PSD matrix) work. In our experiments the number of variables is $p = 200$ and $n = 80$ points are observed. To estimate the true covariance matrix from the noisy observation, first the sample covariance matrix is formed as

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top ,$$

and given as input to various algorithms which provide a new estimate $\hat{\Sigma}$. The methods we compared are the following:

- **Raw sample covariance.** The most basic is to output $\hat{\Sigma}_n$ as the estimate of the covariance, which is not accurate due to presence of noise and underdeterminedness $n < p$.

- **Trace penalty on the PSD cone.** This spectral algorithm solves the following optimization problem in the cone of PSD matrices:

$$\min_{Z \succeq 0} \frac{1}{2} \left\| Z - \hat{\Sigma}_n \right\|_{\mathrm{Fro}}^2 + \lambda \operatorname{Tr} Z .$$

- ℓ_1 **penalty.** In order to approximate the sample covariance $\hat{\Sigma}_n$ by a sparse matrix a basic idea is to soft-threshold it element-by-element. This is equivalent to solving the following convex optimization problem:

$$\min_{Z} \frac{1}{2} \left\| Z - \hat{\Sigma}_n \right\|_{\mathrm{Fro}}^2 + \lambda \|Z\|_1 .$$

- **Trace** + ℓ_1 **penalty.** The restriction of Γ_μ to the PSD cone, which is equivalent to solving the following SDP

$$\min_{Z \succeq 0} \frac{1}{2} \left\| Z - \hat{\Sigma}_n \right\|_{\mathrm{Fro}}^2 + \lambda \Gamma_\mu(Z) .$$

  This approach needs to tune two parameters $\lambda > 0, \mu \in [0,1]$.

- **Sequential sparse PCA.** This is the standard way of estimating multiple sparse principal components which consists of solving the problem for a single component at each step $t = 1 \cdots r$, and deflate to switch to the next $(t+1)$st component. The deflation step used in this algorithm is the orthogonal projection

$$Z_{t+1} = (\mathrm{Id}_p - u_t u_t^\top)\, Z_t\, (\mathrm{Id}_p - u_t u_t^\top) .$$

  The tuning parameters for this approach are the sparsity level k and the number of principal components r.

Sample covariance	Trace	ℓ_1	Trace + ℓ_1	Sequential	$\Omega_{k,\succeq}$
4.20 ± 0.02	0.98 ± 0.01	2.07 ± 0.01	0.96 ± 0.01	0.93 ± 0.08	**0.59 ± 0.03**

Table 3: Relative error of covariance estimation with different methods.

- $\Omega_{k,\succeq}$ **penalty.** The following optimization problem, which is a proximal operator computation, is solved using the active set algorithm:

$$\min_{Z \succeq 0} \frac{1}{2} \left\| Z - \hat{\Sigma}_n \right\|_{\text{Fro}}^2 + \lambda \Omega_{k,\succeq}(Z) \,,$$

  with $\Omega_{k,\succeq}$ the gauge associated with $\mathcal{A}_{k,\succ}$ already introduced in Section 3.4. The two parameters of this method are $\lambda > 0$ and $k \in \mathbb{N} \backslash \{0\}$.

We report the relative errors $\left\| \hat{\Sigma} - \Sigma^\star \right\|_{\text{Fro}} / \left\| \Sigma^\star \right\|_{\text{Fro}}$ over 10 runs of our experiments in Table 3, and a representation of the estimated matrices can be found in Figure 3. We observe that sparse PCA methods using $\Omega_{k,\succeq}$ and also the sequential method using deflation steps outperform spectral and ℓ_1 baselines. In addition, penalizing $\Omega_{k,\succeq}$ is superior to the sequential approach. This was expected since our algorithm minimizes a loss function that is close to the test errors reported, whereas the sequential scheme does not optimize a well-defined objective.

7 Conclusion

In this work, we proposed two new convex penalties, the (k,q)-trace norm and the (k,q)-CUT norm, specifically tailored to the estimation of low-rank matrices with sparse factors. Our motivation for proposing such convex formulations for sparse low-rank matrix inference was twofold. First, it allowed us to consider algorithmic schemes that are better understood when a problem is formulated as a convex optimization problem, even though the complexity of solving the problem exactly remains super-polynomial. Second, using convex geometry allowed us to provide sample complexity and statistical guarantees, and notably to show that the proposed estimators have much better statistical dimension than more standard convex combinations of the ℓ_1 and trace norms. We observed that the improvement exists only for matrices: for sparse vectors, using our penalty (which boils down to the k-support norm in this case) does not improve over the standard ℓ_1 norm, in terms of statistical dimension increase rate.
One limitation of this work is that we assume that the sparsity of the factors is known and fixed. Lifting this constraint and investigating procedures that can adapt to the size of the blocks (like the ℓ_1 norm adapts to the size of the support) is an interesting direction for future research. Another interesting direction is to use the nuclear norm formulation of the (k,q)-trace norm as in Lemma 10 to optimize the regularized problem.

Acknowledgments

We would like to thank Francis Bach for interesting discussions related to this work. This work was supported by the European Research Council (SMAC-ERC-280032).

References

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. Technical Report 1303.6672, arXiv, Mar 2013. URL http://arxiv.org/abs/1303.6672.

A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Stat.*, 37(5B):2877–2921, 2009. URL http://dx.doi.org/10.1214/08-AOS664.

A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, volume 25, pages 1457–1465. Curran Associates, Inc., 2012. URL http://books.nips.cc/papers/files/nips25/NIPS2012_0698.pdf.

F. Bach. Convex relaxations of structured matrix factorizations. Technical Report 1309.3117, arXiv, 2013. URL http://arxiv.org/pdf/1309.3117v1.pdf.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, arXiv, 2008. URL http://arxiv.org/abs/0812.1869.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2011. URL http://dx.doi.org/10.1561/2200000015.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Stat. Sci.*, 27(4):450–468, 2012. URL http://dx.doi.org/10.1214/12-STS394.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In S. Shalev-Shwartz and I. Steinwart, editors, *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Proceedings*, pages 1046–1066. JMLR.org, 2013. URL http://jmlr.org/proceedings/papers/v30/Berthet13.html.

R. Bhatia. *Matrix analysis*. Springer, 1997.

E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, 66(8):1241–1274, 2013. URL http://dx.doi.org/10.1002/cpa.21432.

V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proc. Natl. Acad. Sci. USA*, 110(13):E1181–E1190, Mar 2013. URL http://dx.doi.org/10.1073/pnas.1302293110.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012. URL http://dx.doi.org/10.1007/s10208-012-9135-7.

J. T. Chu. On bounds for the normal integral. *Biometrika*, 42(1/2):263–265, 1954. URL http://dx.doi.org/10.2307/2333443.

A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007. URL http://dx.doi.org/10.1137/050645506.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9:1269–1294, 2008. URL http://jmlr.org/papers/v9/aspremont08a.html.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of the Geometry of Banach Spaces*, volume 1, pages 317 – 366. Elsevier Science B.V., 2001. URL http://dx.doi.org/10.1016/S1874-5849(01)80010-3.

M. M. Deza and M. Laurent. *Geometry of Cuts and Metrics*, volume 15 of *Algorithms and Combinatorics*. Springer Berlin Heidelberg, 1997.

X. V. Doan and S. A. Vavasis. Finding approximately rank-one submatrices with the nuclear norm and ℓ_1 norms. *SIAM J. Optimiz.*, 23(4):2502–2540, 2013. URL http://dx.doi.org/10.1137/100814251.

R. Foygel and L. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Trans. Inform. Theory*, 60(2):1223–1247, 2014. URL http://dx.doi.org/10.1109/TIT.2013.2293654.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. URL http://dx.doi.org/10.1145/1553374.1553431.

G. J. O. Jameson. *Summing and Nuclear Norms in Banach Space Theory*. Number 8 in London Mathematical Society Student Texts. Cambridge University Press, 1987. URL http://dx.doi.org/10.1017/CBO9780511569166.

M. Jerrum. Large cliques elude the Metropolis process. *Random Struct. Alg.*, 3(4):347–359, 1992. URL http://dx.doi.org/10.1002/rsa.3240030402.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11:517–553, 2010. URL http://jmlr.org/papers/volume11/journee10a/journee10a.pdf.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear norm penalization and optimal rates for noisy matrix completion. *Ann. Stat.*, 39(5):2302–2329, 2011. URL http://dx.doi.org/10.1214/11-AOS894.

R. Krauthgamer, B. Nadler, and D. Vilenchik. Do semidefinite relaxations really solve sparse PCA? Technical Report 1306:3690, arXiv, 2013. URL http://arxiv.org/abs/1306.3690.

H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Adv. Neural. Inform. Process Syst.*, volume 19, pages 801–808. MIT Press, 2007. URL http://papers.nips.cc/paper/2979-efficient-sparse-coding-algorithms.

R. Luss and M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Rev.*, 55(1):65–98, 2013. URL http://dx.doi.org/10.1137/110839072.

L. W. Mackey. Deflation methods for sparse PCA. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Adv. Neural. Inform. Process Syst.*, volume 21, pages 1017–1024. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3575-deflation-methods-for-sparse-pca.pdf.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010. URL http://jmlr.csail.mit.edu/papers/v11/mairal10a.html.

B. Moghaddam, Y. Weiss, and Sh. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Neural Information Processing Systems (NIPS)*, volume 18, page 915. MIT Press, 2006.

B. Moghaddam, A. Gruber, Y. Weiss, and S. Avidan. Sparse regression as a sparse eigenvalue problem. In *Information Theory and Applications Workshop, 2008*, pages 121–127. IEEE, Jan 2008. URL http://dx.doi.org/10.1109/ITA.2008.4601036.

S. N Negahban, P. Ravikumar, M. J Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators. *Statistical Science*, 27(4):538–557, 2012.

S. Oymak and B. Hassibi. Sharp mse bounds for proximal denoising. Technical Report 1305.2714, arXiv, 2013. URL http://arxiv.org/abs/1305.2714.

S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. Technical Report 1212.3753, arXiv, 2012. URL http://arxiv.org/abs/1212.3753.

S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized LASSO: A precise analysis. In *51st Annual Allerton Conference on Communication, Control, and Computing, Allerton Park & Retreat Center, Monticello, IL, USA, October 2-4, 2013*, pages 1002–1009. IEEE, 2013. URL http://dx.doi.org/10.1109/Allerton.2013.6736635.

E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low-rank matrices. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/discuss/2012/674.html.

E. Richard, F. Bach, and J.-P. Vert. Intersecting singularities for multi-structured estimation. In S. Dasgupta and D. Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 1157–1165, Atlanta, Georgia, USA, may 2013. JMLR Workshop and Conference Proceedings. URL http://jmlr.org/proceedings/papers/v28/richard13.html.

E. Richard, S. Gaïffas, and N. Vayatis. Link prediction in graphs with autoregressive features. *J. Mach. Learn. Res.*, 15(1):565–593, jan 2014. URL http://jmlr.org/papers/v15/richard14a.html.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, 117(1-2):387–423, 2009. URL http://dx.doi.org/10.1007/s10107-007-0170-0.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutinyok, editors, *Compressed Sensing, Theory and Applications*, pages 210–268. Cambridge University Press, 2012. URL http://dx.doi.org/10.1017/CBO9780511794308.006.

M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009. URL http://dx.doi.org/10.1109/TIT.2009.2032816.

Y.-X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When LRR meets SSC. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Adv. Neural. Inform. Process Syst.*, volume 26, pages 64–72. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4865-provable-subspace-clustering-when-lrr-meets-ssc.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Lin. Alg. Appl.*, 170: 1039–1053, 1992. URL http://dx.doi.org/10.1016/0024-3795(92)90407-2.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, Jul 2009. URL http://dx.doi.org/10.1093/biostatistics/kxp008.

X.-T. Yuan and T. Zhang. Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, 14:889–925, 2013. URL http://www.jmlr.org/papers/volume14/yuan13a/yuan13a.pdf.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J. Comput. Graph. Stat.*, 15(2):265–286, 2006. URL http://dx.doi.org/10.1198/106186006X113430.

A Proofs of results in Sections 2 and 3.

Proof [Proposition 2]
To prove the first claim, note that a matrix of the form $ab^\top$ for $a \in \mathcal{A}_k^{m_1}$ and $b \in \mathcal{A}_q^{m_2}$ has at most kq non-zero terms. Therefore, the decomposition of a matrix with no null entries as a linear combination of such sparse matrices must count at least $\frac{m_1 m_2}{kq}$ terms, which is larger than $m_1 \vee m_2$ when $kq \leq m_1 \wedge m_2$.
To prove second claim, consider the case of the $(2,2)$-SVD for the matrix $Z = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^3$. It is impossible to write Z as the sum of two $(2,2)$-sparse matrices, because it would then have at most 8 non-zero coefficients. But we have the decomposition.

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & \frac{1}{2} & 1 \\ 0 & 1 & 2 \end{pmatrix} - \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix},$$

which shows that the $(2,2)$-rank of Z is 3. Now, given that Z is invariant by any of the 6 permutations of the rows and any of the 6 permutations of the columns, Z admits at least 36 different $(2,2)$-SVDs.
To prove the third claim, observe that the decomposition proposed above for $Z = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^3$ yields 9 left- and right-$(2,2)$-sparse singular vectors that are obviously not orthogonal. It can actually be

shown by systematic enumeration of all possible cases that it is impossible to find any $(2,2)$-sparse-SVD of Z whose left or right singular vectors are orthogonal.

■

Proof [Proposition 6]
To prove the first claim, let us consider the matrix $Z = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^3$. We showed in the proof of Proposition 2 above that its $(2,2)$-rank is equal to 3. We now show that its soft-$(2,2)$-rank is equal to 9. For that purpose, we express any soft-$(2,2)$-SVD of Z as a minimizer of (8), and write the corresponding Lagrangian:

$$\mathcal{L}((Z^{(IJ)})_{I,J}, K) = \sum_{I,J\in\mathcal{G}_2} \left\|Z^{(IJ)}\right\|_* + \mathrm{tr}\Big(K^\top\Big(Z - \sum_{I,J\in\mathcal{G}_2} Z^{(IJ)}\Big)\Big)\,,$$

where $(Z^{(IJ)})_{I,J}$ and K are the primal and dual variables. It is easy to check that the dual solution is the unique subgradient of $\Omega_{2,2}$ at Z which is equal to $K^* = \frac{1}{2}Z$. But any primal solution must satisfy $\mathrm{tr}(K^{*\top} Z^{(IJ)}) = \|Z^{(IJ)}\|_*$. This implies that any primal solution $(Z^{(IJ)})_{I,J}$ satisfies $Z^{(IJ)} \propto \mathbf{1}_I\mathbf{1}_J^\top$. Then, one can check that $((\frac{1}{2})\mathbf{1}_I\mathbf{1}_J^\top)_{I,J\in\mathcal{G}_2}$ forms a basis of $\mathbb{R}^{3\times 3}$ so that any matrix Z admits a unique set of decomposition coefficients on that basis. This proves that the unique solution of (8) is the one such that $Z^{(IJ)} = \frac{1}{4}\mathbf{1}_I\mathbf{1}_J^\top$ for all pairs $(I,J) \in \mathcal{G}_2 \times \mathcal{G}_2$. This unique soft-$(k,q)$-SVD is composed of 9 terms, meaning that the soft-(k,q)-rank of Z is 9 while its (k,q)-rank is 3.
To prove the second claim, let us consider the soft-$(2,2)$-SVDs of $Z = \frac{1}{2}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^4$. By proposition 17, $\frac{1}{2}\|Z\|_1 \leq \Omega_{2,2}(Z)$, but $\frac{1}{2}\|Z\|_1 = 4$ and $2Z = (\mathbf{1}_{\{1,2\}} + \mathbf{1}_{\{3,4\}})(\mathbf{1}_{\{1,2\}} + \mathbf{1}_{\{3,4\}})^\top$ which shows that $\Omega_{2,2}(Z) \leq 4$. So $\Omega_{2,2}(Z) = 4$. Considering that there are 3 ways to partition $\{1,2,3,4\}$ into sets of cardinality 2, Z admits at least 9 different optimal decompositions in the sense of the $(2,2)$-soft-SVD since Z can be written in 9 different ways as the sum of four matrices of $\widetilde{\mathcal{A}}_{2,2}$ with disjoint supports. Each of these decompositions attains the $(2,2)$-rank which is equal to 4. Note also that by convexity any convex combination of these decompositions is also an optimal decomposition in the sense of the soft-$(2,2)$-SVD, but can contain up to 36 terms!
To prove the third claim, let us consider

$$Z_1 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ Z_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix},\ Z = Z_1 + Z_2 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

As Z_1, Z_2, Z are all positive semidefinite we have $\|Z_1\|_* = 2$, $\|Z_2\|_* = 2$, and $\|Z\|_* = 4$. By inequality (7), $\Omega_{2,2}(Z) \geq \|Z\|_* = 4$ which proves that the decomposition $Z = Z_1 + Z_2$ is optimal: $\Omega_{2,2}(Z) = 4$. But $\langle Z_1, Z_2\rangle = 1$. So this decomposition is a decomposition of Z onto linear combination of atoms $\frac{1}{2}Z_1, \frac{1}{2}Z_2 \in \mathcal{A}_{2,2}$ which are not orthogonal. ■

Proof [Lemma 7]
We first show (9) from the definition of the dual norm $\Omega^*_{k,q}$:

$$\begin{aligned}
\Omega^*_{k,q}(Z) &= \max_K \{\langle K, Z\rangle \ :\ \Omega_{k,q}(K) \leq 1\} \\
&= \max_{a,b} \{\langle Z, ab^\top\rangle \ :\ ab^\top \in \mathcal{A}_{k,q}\} \\
&= \max_{a,b} \{a^\top Z b \ :\ \|a\|_0 \leq k\ ,\ \|b\|_0 \leq q\ ,\ \|a\|_2 = \|b\|_2 = 1\} \\
&= \max_{I,J} \left\{\|Z_{I,J}\|_{\mathrm{op}} \ :\ I \in \mathcal{G}_k^{m_1}\ ,\ J \in \mathcal{G}_q^{m_2}\right\},
\end{aligned}$$

where the second equality follows from the fact that the maximization of a linear form over a bounded convex set is attained at one of the extreme points of the set. Given this closed-form expression of the dual norm, we prove the variational formulation (8) for the primal norm $\Omega_{k,q}$. Consider the function $\check{\Omega}_{k,q}$ defined by

$$\check{\Omega}_{k,q}(Z) = \inf \left\{ \sum_{(I,J)\in\mathcal{G}_k^{m_1}\times\mathcal{G}_q^{m_2}} \left\| Z^{(I,J)} \right\|_* \ : \ Z = \sum_{(I,J)} Z^{(I,J)} \ , \ \mathrm{supp}(Z^{(I,J)}) \subset I \times J \right\} .$$

Since $\check{\Omega}_{k,q}(Z)$ is defined as the infimum of a jointly convex function of Z and $(Z^{(I,J)})_{I\in\mathcal{G}_k^{m_1},\, J\in\mathcal{G}_q^{m_2}}$ obtained by minimizing w.r.t. to the latter variables, it is a an elementary fact from convex analysis that $\check{\Omega}_{k,q}$ is a convex function of Z. It is also symmetric and positively homogeneous, which together with convexity prove that $\check{\Omega}_{k,q}$ defines a norm. We can compute its dual norm as

$$\begin{aligned}
\check{\Omega}_{k,q}{}^*(K) &= \max_Z \left\{ \langle K, Z\rangle \ : \ \check{\Omega}_{k,q}(Z) \le 1 \right\} \\
&= \max_{(Z^{(IJ)})_{(I,J)}} \left\{ \langle K, \sum_{(I,J)} Z^{(IJ)} \rangle \ : \ \sum_{(I,J)} \left\| Z^{(IJ)} \right\|_* \le 1 \ , \ \mathrm{supp}(Z^{(IJ)}) \subset I \times J \right\} \\
&= \max_{(Z^{(IJ)})_{(I,J)}, (\eta^{(IJ)})_{(I,J)}} \left\{ \sum_{(I,J)} \eta^{(I,J)} \langle K_{I,J}, Z^{(IJ)} \rangle \ : \ \left\| Z^{(IJ)} \right\|_* \le \eta^{(IJ)}, \quad \sum_{(I,J)} \eta^{(IJ)} \le 1 \right\} \\
&= \max_{(\eta^{(IJ)})_{(I,J)}} \left\{ \sum_{(I,J)} \eta^{(IJ)} \left\| K_{I,J} \right\|_{\mathrm{op}} \ : \ \sum_{(I,J)} \eta^{(IJ)} \le 1 \right\} \\
&= \max_{(I,J)} \left\| K_{I,J} \right\|_{\mathrm{op}} \\
&= \Omega^*_{k,q}(K) .
\end{aligned}$$

This proves that $\Omega_{k,q}(K) = \check{\Omega}_{k,q}(K)$ since a norm is uniquely characterized by its dual norm. Finally, to show (10) we use the general characterization of the subdifferential of a norm (e.g., Watson, 1992):

$$G \in \partial\Omega_{k,q}(A) \Leftrightarrow \begin{cases} \Omega_{k,q}(A) = \langle G, A\rangle , \\ \Omega^*_{k,q}(G) \le 1 . \end{cases}$$

Let us denote a subgradient by $G = A + Z$. Since $A = ab^\top$ is an atom, we have $\Omega_{k,q}(A) = 1$. In addition, $\|A\|^2_{\mathrm{Fro}} = \mathrm{Tr}(ba^\top ab^\top) = 1$, therefore the condition $\Omega_{k,q}(A) = \langle G, A\rangle$ boils down to $\langle Z, A\rangle = 0$. Given the characterization of the dual norm (9), we therefore get:

$$\partial\Omega_{k,q}(A) = \left\{ A + Z \ : \ \langle A, Z\rangle = 0, \ \forall (I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2} \ \ \left\| A_{I,J} + Z_{I,J} \right\|_{\mathrm{op}} \le 1 \right\} .$$

Let now

$$\mathcal{D}(A) = \left\{ A + Z \ : \ AZ^\top_{I_0,J_0} = 0, \ A^\top Z_{I_0,J_0} = 0, \ \forall (I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2} \ \ \left\| A_{I,J} + Z_{I,J} \right\|_{\mathrm{op}} \le 1 \right\} .$$

Since $\langle A, Z\rangle = \langle A, Z_{I_0,J_0}\rangle = \mathrm{Tr}\,(A^\top Z_{I_0,J_0})$, it is clear that $\mathcal{D}(A) \subset \partial\Omega_{k,q}(A)$. Conversely, let $G = A+Z \in \partial\Omega_{k,q}(A)$. Then $\langle A, Z\rangle = \langle ab^\top, Z_{I_0,J_0}\rangle = a^\top Z_{I_0,J_0} b = 0$, and therefore, by Pythagorean equality applied to the orthogonal vectors a and $Z_{I_0,J_0}b$:

$$\left\| (A_{I_0,J_0} + Z_{I_0,J_0})\, b \right\|_2^2 = \left\| ab^\top b + Z_{I_0,J_0} b \right\|_2^2 = \left\| a + Z_{I_0,J_0} b \right\|_2^2 = 1 + \left\| Z_{I_0,J_0} b \right\|_2^2 ,$$

but since $\|A_{I_0,J_0} + Z_{I_0,J_0}\|_{\text{op}} \leq 1$ and $\|b\|_2 = 1$ we must have $\|Z_{I_0,J_0} b\|_2 = 0$. This shows that $AZ_{I_0,J_0}^\top = ab^\top Z_{I_0,J_0}^\top = 0$. The same reasoning starting with the orthogonal vectors b and $Z_{I_0,J_0}^\top a$ shows that we also have $A^\top Z_{I_0,J_0} = 0$, implying that $\partial\Omega_{k,q}(A) \subset \mathcal{D}(A)$. This concludes the proof that $\partial\Omega_{k,q}(A) = \mathcal{D}(A)$, as claimed in (10). ■

Proof [Lemma 10]
Let ν be the nuclear norm induced by two atomic norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, induced themselves respectively by the two atom sets $\mathcal{A}_1$ and $\mathcal{A}_2$. Let $\mathcal{A} = \left\{ab^\top \ : \ a \in \mathcal{A}_1\,,\, b \in \mathcal{A}_2\right\}$ and $B = \text{Conv}\big(\mathcal{A}\big)$, then the key argument is to note that we have

$$\left\{ab^\top \ : \ \|a\|_\alpha \leq 1,\ \ \|b\|_\beta \leq 1\right\} \subset B\,.$$

Indeed, if $a = \sum_i \lambda_i a_i$ and $b = \sum_j \lambda'_j b_j$ with $a_i \in \mathcal{A}_1$, $b_j \in \mathcal{A}_2$ and $\sum_i \lambda_i = \sum_j \lambda'_j = 1$, then with $\mu_{ij} := \lambda_i \lambda'_j$, we have $ab^\top = \sum_{i,j} \mu_{ij} a_i b_j^\top$ and $\sum_{i,j} \mu_{ij} = 1$. The inclusion is then proved by density. By (12) the dual norm of ν satisfies

$$\nu^*(Z) = \sup\left\{a^\top Z b \ : \ \|a\|_\alpha \leq 1\ ,\ \ \|b\|_\beta \leq 1\right\}\,,$$

so that

$$\nu^*(Z) \leq \sup\left\{\langle Z, ab^\top\rangle \ : \ ab^\top \in B\right\} = \sup\left\{\langle Z, ab^\top\rangle \ : \ ab^\top \in \mathcal{A}\right\} \leq \nu^*(Z)\,,$$

where the middle equality is due to the fact that the maximum of a linear function on a convex set is attained at a vertex. We therefore have $\nu^*(Z) = \sup\left\{\langle Z, A\rangle \ : \ A \in \mathcal{A}\right\}$. Given (4), this shows that ν is the atomic norm induced by $\mathcal{A}$. ■

Proof [Theorem 11]
Since the (k,q)-trace norm is the atomic norm induced by the atom set (5), Lemma 10 tells us that it is also the nuclear norm induced by the two atomic norms with atom sets $\mathcal{A}_k^{m_1}$ and $\mathcal{A}_q^{m_2}$, which correspond exactly to the so-called k- and q-support norms of Argyriou et al. (2012).
To prove the second statement, we proceed similarly to get that the (k,q)-CUT norm is the nuclear norm induced by the two atomic norms with atom sets $\widetilde{\mathcal{A}}_k^{m_1}$ and $\widetilde{\mathcal{A}}_k^{m_2}$. Calling κ_k and κ_q these norms, we obtain an explicit formulation as follows:

$$\begin{aligned}
\kappa_k(w) &= \max_s \left\{\langle s, w\rangle \ : \ \kappa_k^*(s) \leq 1\right\} \\
&= \max\left\{\langle s, w\rangle \ : \ \frac{1}{\sqrt{k}} \sum_{i=1}^k |s_{(i)}| \leq 1\right\} \\
&= \begin{cases} \dfrac{1}{k\sqrt{k}} \|w\|_1 & \text{if} \quad \|w\|_1 \geq k \|w\|_\infty \\ \dfrac{1}{\sqrt{k}} \|w\|_\infty & \text{if} \quad \|w\|_1 \leq k \|w\|_\infty \end{cases} \\
&= \frac{1}{\sqrt{k}} \max\left(\|w\|_\infty, \frac{1}{k}\|w\|_1\right)\,.
\end{aligned}$$

■

Proof [Lemma 12]

The form of θ_k^* follows immediately from the fact that $\theta_k^*(w) = \max\{a^\top w \ : \ a \in \mathcal{A}_k\}$. Similarly for κ_k^*, we have

$$\kappa_k^*(s) = \max\left\{\langle a, s\rangle \ : \ a \in \widetilde{\mathcal{A}}_k\right\} = \max_{I:|I|=k} \|s_I\|_1 = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} |s_{(i)}|,$$

where $s_{(i)}$ denotes the the ith largest element of s in absolute value. This norm is proportional to a norm known as the vector k-norm or 1-k symmetric norm gauge. ■

Proof [Proposition 13]
To prove the first claim, we show a counterexample for the $(2,2)$-SVD in $\mathbb{R}^{4\times 4}$. Let $I = \{1,2\}$ and $J = \{3,4\}$. The matrix $Z = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^4$ can be written as $Z = Z_{I,I} + Z_{I,J} + Z_{J,I} + Z_{J,J}$, so its $(2,2)$-rank is less than 4. However, it is not possible to write it as a sum of less than 6 symmetric $(2,2)$-sparse matrices, because each of these matrices can only make one coefficient above the non-diagonal non-zero. Therefore, its $(2,2)$-SVD must contain non-symmetric terms.
To prove the second claim, note first that the case $k = 1$ is peculiar and not representative of the general case because the span of the PSD matrices of sparsity 1 are only the diagonal matrices, while the span of rank one PSD matrices of sparsity $k \times k$ for $k > 1$ is all the symmetric matrices. Now, we claim that it is not possible to write $Z = \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^3$ as a sum of PSD matrices that are $(2,2)$-sparse and PSD. Indeed, if this was the case, this would imply the existence of a non zero vector v with a support of size at most 2 such that $Z - vv^\top \succ 0$. Since the only eigenvector of Z associated with a non-zero eigenvalue is the constant vector this is impossible.

■

B Proofs of results in Section 4.1

Proof [Lemma 14]
We prove a more general result than Lemma 14. Let $\Omega : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}$ be any matrix norm, and $\mathcal{X} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}^n$ be a linear map. We denote by X_i $(i = 1,\dots,n)$ the i-th design matrix defined by $\mathcal{X}(Z)_i = \langle Z, X_i\rangle$. For a given matrix $Z^\star \in \mathbb{R}^{m_1\times m_2}$, assume we observe:

$$Y = \mathcal{X}(Z^\star) + \epsilon\,, \tag{41}$$

where $\epsilon \in \mathbb{R}^n$ is a centered random noise vector. We consider the following estimator of $Z^\star$:

$$\hat{Z}_\Omega \in \arg\min_Z \frac{1}{2n}\|Y - \mathcal{X}(Z)\|_2^2 + \lambda\Omega(Z)\,, \tag{42}$$

for some value of the parameter $\lambda > 0$. The following result generalizes standard results known for the ℓ_1 and trace norms (e.g., Koltchinskii et al., 2011, Theorem 1) to any norm Ω.

Theorem 27 *If* $\lambda \geq \frac{1}{n}\Omega^*(\sum_{i=1}^n \epsilon_i X_i)$ *then*

$$\frac{1}{2n}\left\|\mathcal{X}(\hat{Z}_\Omega - Z^\star)\right\|_2^2 \leq \inf_Z\left\{\frac{1}{2n}\|\mathcal{X}(Z - Z^\star)\|_2^2 + 2\lambda\Omega(Z)\right\}. \tag{43}$$

Lemma 14 is then a simple consequence of Theorem 27 by taking for $\mathcal{X}$ the identity map, upper bounding the right-hand side of (43) by the value $2\lambda\Omega(Z^\star)$ it takes for $Z = Z^\star$, and replacing λ by λ/n. ■

Proof [Theorem 27]
By definition of $\hat{Z}_\Omega$ (42), we have for all Z:

$$\frac{1}{2n}\left\|Y - \mathcal{X}(\hat{Z}_\Omega)\right\|_2^2 \leq \frac{1}{2n}\|Y - \mathcal{X}(Z)\|_2^2 + \lambda\left(\Omega(Z) - \Omega(\hat{Z}_\Omega)\right),$$

which after developing the squared norm and replacing Y by (41) gives

$$\frac{1}{2n}\left\|\mathcal{X}(\hat{Z}_\Omega)\right\|_2^2 - \frac{1}{n}\langle\mathcal{X}(Z^\star) + \epsilon, \mathcal{X}(\hat{Z}_\Omega)\rangle \leq \frac{1}{2n}\|\mathcal{X}(Z)\|_2^2 - \frac{1}{n}\langle\mathcal{X}(Z^\star) + \epsilon, \mathcal{X}(Z)\rangle + \lambda\left(\Omega(Z) - \Omega(\hat{Z}_\Omega)\right),$$

and therefore

$$\frac{1}{2n}\left\|\mathcal{X}(\hat{Z}_\Omega - Z^\star)\right\|_2^2 \leq \frac{1}{2n}\|\mathcal{X}(Z - Z^\star)\|_2^2 + \frac{1}{n}\langle\epsilon, \mathcal{X}(\hat{Z}_\Omega - Z)\rangle + \lambda\left(\Omega(Z) - \Omega(\hat{Z}_\Omega)\right). \tag{44}$$

Now, using the fact (true for any norm) that $\Omega(A)\Omega^\star(B) \geq \langle A, B\rangle$ for any vectors $A, B \in \mathbb{R}^n$, and taking $\lambda \geq \frac{1}{n}\Omega^*(\sum_{i=1}^n \epsilon_i X_i)$, we can upper bound the second term of the right-hand side of (44) by:

$$\begin{aligned}
\frac{1}{n}\langle\epsilon, \mathcal{X}(\hat{Z}_\Omega - Z)\rangle &= \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathcal{X}(\hat{Z}_\Omega - Z)_i \\
&= \frac{1}{n}\sum_{i=1}^n \epsilon_i \langle X_i, \hat{Z}_\Omega - Z\rangle \\
&= \frac{1}{n}\langle\sum_{i=1}^n \epsilon_i X_i, \hat{Z}_\Omega - Z\rangle \\
&\leq \frac{1}{n}\Omega^\star\left(\sum_{i=1}^n \epsilon_i X_i\right)\Omega\left(\hat{Z}_\Omega - Z\right) \\
&\leq \lambda\Omega\left(\hat{Z}_\Omega - Z\right)
\end{aligned}$$

Plugging this bound back in (44) finally gives

$$\begin{aligned}
\frac{1}{2n}\left\|\mathcal{X}(\hat{Z}_\Omega - Z^\star)\right\|_2^2 &\leq \frac{1}{2n}\|\mathcal{X}(Z - Z^\star)\|_2^2 + \lambda\Omega(\hat{Z}_\Omega - Z) + \lambda\left(\Omega(Z) - \Omega(\hat{Z}_\Omega)\right) \\
&\leq \frac{1}{2n}\|\mathcal{X}(Z - Z^\star)\|_2^2 + 2\lambda\Omega(Z),
\end{aligned}$$

the last inequality being due to the triangle inequality. ■

Before proving Propositon 15, let us first derive an intermediary results useful to obtain an upper bound on the dual (k, q)-trace norm of a random matrix with i.i.d. normal entries.

Lemma 28 *Let G be a $m_1 \times m_2$ random matrix with i.i.d. normally distributed entries. Then*

$$\mathbb{E}\max_{I\in\mathcal{G}_k, J\in\mathcal{G}_q}\|G_{I,J}\|_{\text{op}}^2 \leq 16\left[\left(k\log\frac{m_1}{k} + q\log\frac{m_2}{q}\right) + 2(k+q)\right].$$

Proof [Lemma 28]
For a random matrix $H \in \mathbb{R}^{k\times q}$ with i.i.d. standard normal entries, we have the following concentration inequality (e.g., Davidson and Szarek, 2001): for $s \geq 0$,

$$\mathbb{P}[\|H\|_{\text{op}} > \sqrt{k} + \sqrt{q} + s] \leq \exp(-s^2/2)\,. \tag{45}$$

Denoting $R = 2\left(\sqrt{k} + \sqrt{q}\right)$, and $f(x) = e^{tx^2}$, we have the sequence of inequalities

$$\begin{aligned}
\mathbb{E}\exp(t\,\|H\|_{\text{op}}^2) &= \mathbb{E}f(\|H\|_{\text{op}}) \\
&= \int_1^\infty \mathbb{P}[f(\|H\|_{\text{op}}) > h]\ \ dh \\
&\leq \int_1^{1+f(R)} 1\ dh + \int_{1+f(R)}^\infty \mathbb{P}[f(\|H\|_{\text{op}}) > h]dh \\
&= f(R) + \int_0^\infty \mathbb{P}[\|H\|_{\text{op}} > f^{-1}(f(R) + 1 + \zeta)]d\zeta \\
&\leq f(R) + \int_0^\infty \mathbb{P}[\|H\|_{\text{op}} > \frac{1}{2}R + \frac{1}{2}f^{-1}(1+\zeta)]d\zeta \qquad (46) \\
&\leq f(R) + \int_0^\infty 8ts\ \exp\left(-s^2/2 + 4ts^2\right) ds \qquad (47) \\
&\leq f(R) + 4\frac{t}{\frac{1}{2} - 4t} \qquad (48) \\
&\leq \exp(8t(k+q)) + \frac{8t}{1-8t}\,,
\end{aligned}$$

where the change of variable used in (47) is $1 + \zeta = f(2s) = e^{4ts^2}$, (48) is true for any $t < \frac{1}{8}$, and (46) follows from the property of the inverse $f^{-1}(z) = \sqrt{\frac{\log(z)}{t}}$ that it is strictly increasing on $[1;\infty)$ and sandwiched via

$$\frac{1}{2}\left\{f^{-1}(z) + f^{-1}(z')\right\} \leq f^{-1}(z+z') \leq f^{-1}(z) + f^{-1}(z')\,. \tag{49}$$

Take now $t = \frac{1}{8} - \frac{1}{8(k+q)}$. Since $k+q \geq 2$, we have $1/16 \leq t < 1/8$. Therefore,

$$\begin{aligned}
\mathbb{E} \max_{I,J} \|G_{I,J}\|_{\mathrm{op}}^2 &= \frac{1}{t} \log \left\{ \exp t\, \mathbb{E} \max_{I,J} \|G_{I,J}\|_{\mathrm{op}}^2 \right\} \\
&\leq \frac{1}{t} \log \left\{ \mathbb{E} \exp(t \max_{I,J} \|G_{I,J}\|_{\mathrm{op}}^2) \right\} \\
&\leq \frac{1}{t} \log \left\{ \sum_{I,J} \mathbb{E} \exp(t \|G_{I,J}\|_{\mathrm{op}}^2) \right\} \\
&\leq \frac{1}{t} \log \left\{ \binom{m_1}{k} \binom{m_2}{q} \mathbb{E} \exp(t \|H\|_{\mathrm{op}}^2) \right\} \\
&\leq \frac{1}{t} \log \left\{ \left(\frac{e\ m_1}{k}\right)^k \left(\frac{e\ m_2}{q}\right)^q \left(e^{8t(k+q)} + \frac{8t}{1-8t} \right) \right\} \\
&= \frac{1}{t} \left[\left(k \log \frac{m_1}{k} + q \log \frac{m_2}{q} \right) + k + q + 8t(k+q) + \log \left(1 + \frac{8t}{1-8t} e^{-8t(k+q)} \right) \right] \\
&\leq 16 \left[\left(k \log \frac{m_1}{k} + q \log \frac{m_2}{q} \right) + k + q \right] + 8(k+q) + \frac{8}{1-8t} e^{-8t(k+q)} \\
&\leq 16 \left[\left(k \log \frac{m_1}{k} + q \log \frac{m_2}{q} \right) + 2(k+q) \right] ,
\end{aligned}$$

where in the last inequality we simply used $8/(1-8t) = 8(k+q)$ and $\exp(-8t(k+q)) \leq 1$. ■

Proof [Propositon 15]
From Lemma 28 we have:

$$\begin{aligned}
\mathbb{E}\, \Omega_{k,q}^*(G) &= \mathbb{E} \max_{I \in \mathcal{G}_k, J \in \mathcal{G}_q} \|G_{I,J}\|_{\mathrm{op}} \\
&\leq \left(\mathbb{E} \max_{I \in \mathcal{G}_k, J \in \mathcal{G}_q} \|G_{I,J}\|_{\mathrm{op}}^2 \right)^{\frac{1}{2}} \\
&\leq 4 \left[\left(k \log \frac{m_1}{k} + q \log \frac{m_2}{q} \right) + 2(k+q) \right]^{\frac{1}{2}} \\
&\leq 4 \left(\sqrt{k \log \frac{m_1}{k} + 2k} + \sqrt{q \log \frac{m_2}{q} + 2q} \right)
\end{aligned}$$

The upper bounds for the ℓ_1 and trace norms are standard. See Vershynin (2012, Theorem. 5.32) for the tight upper bound on the operator norm $\mathbb{E}\, \|G\|_{\mathrm{op}} \leq \sqrt{m_1} + \sqrt{m_2}$, and for the upper bound on the element-wise ℓ_∞ norm of G, use Jensen inequality followed by upper bounding the maximum of nonnegative scalars by their sum:

$$\begin{aligned}
\exp\left(t\ \mathbb{E}\ \|G\|_\infty\right) &\leq \mathbb{E} \exp\left(t\ \ \|G\|_\infty\right) \\
&\leq m_1 m_2 \exp(t^2/2)\,.
\end{aligned}$$

Taking $t = \sqrt{2 \log(m_1 m_2)}$ in the logarithms of the last inequality gives $\mathbb{E}\ \|G\|_\infty \leq \sqrt{2 m_1 m_2}$. ■

C Some cone inclusions (Proofs of results in Section 4.2.2)

Let us start with a simple result useful to prove inclusions of tangent cones.

Lemma 29 *Let f and g two convex functions from $\mathbb{R}^d$ such that $f \leq g$ and let x^* such that $f(x^*) = g(x^*)$. Then $T_g(x^*) \subset T_f(x^*)$.*

Proof [Lemma 29]
Let $h \in \mathbb{R}^d$ and $\tau > 0$ such that $g(x^* + \tau h) \leq g(x^*)$. Then we also have

$$f(x^* + \tau h) \leq g(x^* + \tau h) \leq g(x^*) = f(x^*)\,,$$

and therefore, for any $\tau > 0$,

$$\left\{h \in \mathbb{R}^d \;:\; g(x^* + \tau h) \leq g(x^*)\right\} \subset \left\{h \in \mathbb{R}^d \;:\; f(x^* + \tau h) \leq f(x^*)\right\}.$$

From the definition (26) of the tangent cone we deduce, by taking the union over $\tau > 0$ and the closure of this inclusion, that $T_g(x^*) \subset T_f(x^*)$. ■

We can now prove the results in Section 4.2.2
Proof [Proposition 17]
Consider a matrix $A = ab^\top \in \widetilde{\mathcal{A}}_{k,q}$. We have $\|A\|_* = \|a\|_2 \|b\|_2 = 1$, and $\|A\|_1 = \|a\|_1 \|b\|_1 = \sqrt{kq}$. Since A is an atom of both the norm $\Omega_{k,q}$ and the norm $\widetilde{\Omega}_{k,q}$ we have $\Omega_{k,q}(A) = \widetilde{\Omega}_{k,q}(A) = 1$ so that, for any $\mu \in [0,1]$,

$$\Gamma_\mu(A) = \|A\|_* = \frac{1}{\sqrt{kq}} \|A\|_1 = \Omega_{k,q}(A) = \widetilde{\Omega}_{k,q}(A) = 1\,.$$

Besides, for any matrix $K \in \mathbb{R}^{m_1 \times m_2}$, for all $(I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}$, we have $\|K_{I,J}\|_{\text{op}} \leq \|K\|_{\text{op}}$ and $\|K_{I,J}\|_{\text{op}} \leq \|K_{I,J}\|_{\text{Fro}} \leq \sqrt{kq}\,\|K_{I,J}\|_\infty$ so that $\Omega^*_{k,q}(K) \leq \|K\|_{\text{op}}$ and $\Omega^*_{k,q}(K) \leq \sqrt{kq}\max_{I,J}\|K_{I,J}\|_\infty = \sqrt{kq}\,\|K\|_\infty$. Given that $\widetilde{\mathcal{A}}_{k,q} \subset \mathcal{A}_{k,q}$, we also have that

$$\widetilde{\Omega}^*_{k,q}(K) = \max_{A \in \widetilde{\mathcal{A}}_{k,q}} \langle A, K\rangle \leq \max_{A \in \mathcal{A}_{k,q}} \langle A, K\rangle = \Omega^*_{k,q}(K)\,.$$

By Fenchel duality, we therefore have for any $Z \in \mathbb{R}^{m_1 \times m_2}$ and $\mu \in [0,1]$:

$$\frac{\mu}{\sqrt{kq}}\|Z\|_1 + (1-\mu)\|Z\|_* \leq \Omega_{k,q}(Z) \leq \widetilde{\Omega}_{k,q}(Z)\,.$$

■

Proof [Corollary 18]
Combining Proposition 17 with Lemma 29 directly gives (31). (32) is then a direct consequence of the definition of the statistical dimension (27). ■

Proof [Corollary 19]
A necessary and sufficient condition for exact recovery is the so called null space property which is the event that $T_\Omega(Z^*) \cap \text{Ker}(\mathcal{X}) = \{0\}$, where $\text{Ker}(\mathcal{X})$ is the kernel of the linear transformation $\mathcal{X}$ (Chandrasekaran et al., 2012, Proposition 2.1). The result therefore follows from the inclusion of the cones stated in Corollary 18. ■

Proof [Proposition 20]

Let $a \in \widetilde{\mathcal{A}}_k^m$ with $\mathrm{supp}(a) = I_0$, meaning that $|a_i| = 1/\sqrt{k}$ for $i \in I_0$ and $a_i = 0$ for $i \in I_0^\complement$. The sub differential of the scaled ℓ_1 norm Γ_1 at a is

$$\partial\Gamma_1(a) = \left\{ s \in \mathbb{R}^m \ : \ s_i = \mathrm{sign}(a_i) \text{ for } i \in I_0 \,, \ |s_i| \leq 1 \text{ for } i \in I_0^\complement \right\} .$$

From (10), we get that the subdifferential of θ_k at a is

$$\partial\theta_k(a) = \{ a + z \ : \ \forall i \,, a_i z_i = 0 \text{ and } \forall I \in \mathcal{G}_k^m \,, \|a_I + z_I\| \leq 1 \} .$$

The first condition is equivalent to $z_i = 0$ for $i \in I_0$, which implies that the second is equivalent to $|z_i| \leq 1/\sqrt{k}$ for $i \in I_0^\complement$. We deduce that $s = a + z \in \partial\theta_k(a)$ if and only if $s_i = a_i$ for $i \in I_0$ and $|s_i| \leq 1/\sqrt{k}$ for $i \in I_0^\complement$, *i.e.*,

$$\partial\theta_k(a) = \frac{1}{\sqrt{k}} \partial\Gamma_1(a) \,.$$

This shows that the subdifferentials of Γ_1 and θ_k have the same conic hull, and Proposition 20 follows by noting that the tangent cone is the polar cone of the conic hull of the subdifferential (Rockafellar, 1997, Theorem 23.7). ■

D Upper bound on the statistical dimension of $\Omega_{k,q}$ (proof of Proposition 23)

The aim of this appendix is to prove the upper bound on the statistical dimension $\Omega_{k,q}$ given in Proposition 23. Given its level of technicality, we split the proof in several parts. We start with preliminaries and notations in Section D.1, before proving Proposition 23 in Section D.2. The proofs of several technical results needed in Section D.2 are postponed to Section D.3, D.4 and D.5.

D.1 Preliminaries and notations

Let us start with some notations used throughout Appendix D. $A = ab^\top \in \mathcal{A}_{k,q}$ is an atom of $\Omega_{k,q}$, with $I_0 = \mathrm{supp}(a)$ and $J_0 = \mathrm{supp}(b)$. $\gamma = \gamma(a,b)$ refers to the atom strength of A (Definition 22). For any $I \in \mathcal{G}_k^{m_1}$ and $J \in \mathcal{G}_q^{m_2}$, let $u_I = a_I / \|a_I\|_2$ and $v_J = b_J / \|b_J\|_2$. Note that while a_I is a subvector of a, the notation u_I does not refer to a subvector of some vector u and that therefore $[u_I]_{I_0} \neq [u_{I_0}]_I = a_I$ since $\|a_{I_0}\| = \|a\| = 1$.

To analyze the statistical dimension (27) of $\Omega_{k,q}$ at A, it is useful to express it as follows (Chandrasekaran et al., 2012, Proposition 3.6):

$$\mathfrak{S}(A, \Omega_{k,q}) := \mathbb{E}\left[\mathrm{dist}\left(G, N_{\Omega_{k,q}}(A)\right)^2 \right] , \tag{50}$$

where $N_{\Omega_{k,q}}(A)$ is the normal cone of $\Omega_{k,q}$ at A (*i.e.*, the conic hull of the subdifferential of $\Omega_{k,q}$ at A) and $\mathrm{dist}\left(G, N_{\Omega_{k,q}}(A)\right)$ denotes the Frobenius distance of the Gaussian matrix G with i.i.d. standard normal entries to $N_{\Omega_{k,q}}(A)$. In order to upper bound this quantity, it is therefore important to characterize precisely the normal cone $N_{\Omega_{k,q}}(A)$.

For that purpose, let us introduce further notations. We consider the following subspace of $\mathbb{R}^{m_1 \times m_2}$

$$\mathrm{span}(A) = \left\{ LA + AR \ : \ L \in \mathbb{R}^{m_1 \times m_1}, \ R \in \mathbb{R}^{m_2 \times m_2} \right\} ,$$

and denote by $\mathcal{P}_A$ and $\mathcal{P}_A^\perp$ the orthogonal projectors onto $\mathrm{span}(A)$ and $\mathrm{span}^\perp(A)$ respectively. Since $A = ab^\top$ with $\|a\|_2 = \|b\|_2 = 1$, we have the closed-form expressions $\mathcal{P}_A^\perp(Z) = (Id_{m_1} - aa^\top) Z (Id_{m_2} - bb^\top)$.

For any $(I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}$, consider now the subspace

$$\mathrm{span}_{I,J}(A) = \left\{ L_{I,I} A_{I,J} + A_{I,J} R_{J,J} \ : \ L \in \mathbb{R}^{m_1 \times m_1},\ R \in \mathbb{R}^{m_2 \times m_2} \right\},$$

and its orthogonal

$$\mathrm{span}_{I,J}^{\perp}(A) = \left\{ Z \in \mathbb{R}^{m_1 \times m_2} \ : \ A_{I,J} Z_{I,J}^{\top} = A_{I,J}^{\top} Z_{I,J} = 0 \right\}.$$

Note that $\mathrm{span}_{I_0,J_0}^{\perp}(A)$ is related to the subdifferential of $\Omega_{k,q}$ at A, since according to (10) we can write it as

$$\partial \Omega_{k,q}(A) = \left\{ A + Z \ : \ Z \in \mathrm{span}_{I_0,J_0}^{\perp}(A)\,,\ \forall (I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2} \ \ \|A_{I,J} + Z_{I,J}\|_{\mathrm{op}} \leq 1 \right\}. \tag{51}$$

It is possible to estimate the dimension of $\mathrm{span}_{I_0,J_0}^{\perp}(A)$ as follows:

Lemma 30 *The dimension of $\mathrm{span}_{I_0,J_0}(A)$ is $k+q-1$.*

Proof [Lemma 30]
For $A = ab^{\top}$, the range of $L \mapsto L_{I_0,I_0} A_{I_0,J_0}$ equals the range of $\alpha_{I_0} \mapsto \alpha_{I_0} b^{\top}$ which has dimension $|I_0| = k$. By the same token, the range of $R \mapsto A_{I_0,J_0} R_{J_0,J_0}$ has dimension q. By definition of $\mathrm{span}_{I_0,J_0}(A)$ we therefore have

$$\mathrm{span}_{I_0,J_0}(A) = \left\{ \alpha_{I_0} b^{\top} + a \beta_{J_0}^{\top} \ : \ \alpha \in \mathbb{R}^{m_1}, \beta \in \mathbb{R}^{m_2} \right\}$$

and therefore by the inclusion-exclusion principle $\dim\left(\mathrm{span}_{I_0,J_0}(A)\right) = k+q-1$. ■

Finally we denote by $\Pi_{A,I,J}$ the projector onto $\mathrm{span}_{I,J}(A)$, and by $\Pi_{A,I,J}^{\perp}$ the projector onto $\mathrm{span}_{I,J}^{\perp}(A)$. They satisfy respectively

$$\Pi_{A,I,J}(Z) = \mathcal{P}_{A_{I,J}}(Z_{I,J}) \quad \text{and} \quad \Pi_{A,I,J}^{\perp}(Z) = Z - \Pi_{A,I,J}(Z) = Z - \mathcal{P}_{A_{I,J}}(Z_{I,J})\,.$$

D.2 Proof of Proposition 23

Proof [Proposition 23]
In order to upper bound the statistical dimension of $\Omega_{k,q}$ at A, we associate to any matrix G a matrix $\Xi(G)$ belonging to the normal cone $N_{\Omega_{k,q}}(A)$, where $\Xi : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{m_1 \times m_2}$ is measurable. From the characterization of the statistical dimension (50), since $\mathrm{dist}\left(G, N_{\Omega_{k,q}}(A)\right) \leq \|G - \Xi(G)\|_{\mathrm{Fro}}$, we will then get the upper bound:

$$\mathfrak{S}(A, \Omega_{k,q}) = \mathbb{E}\left[\mathrm{dist}\left(G, N_{\Omega_{k,q}}(A)\right)^2 \right] \leq \mathbb{E}\,\|G - \Xi(G)\|_{\mathrm{Fro}}^2\,. \tag{52}$$

The main steps in the proof are then (i) to define the mapping Ξ, (ii) to show that $\Xi(G) \in N_{\Omega_{k,q}}(A)$ for all G, and (iiii) to upper bound $\mathbb{E}\,\|G - \Xi(G)\|_{\mathrm{Fro}}^2$ in order to derive an upper bound on $\mathfrak{S}(A, \Omega_{k,q})$ by (52).
Given a measurable function $\epsilon : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}$, let us therefore consider the mapping Ξ:

$$\forall G \in \mathbb{R}^{m_1 \times m_2}, \quad \Xi(G) := \epsilon(G) A + \Pi_{A,I_0,J_0}^{\perp}(G)\,. \tag{53}$$

The following lemma provides a mapping ϵ to ensure that $\Xi(G) \in N_{\Omega_{k,q}}(A)$.

Lemma 31 *Let $\epsilon(G)^2$ be equal to*

$$\frac{16}{\gamma^2}\left\|G_{I_0,J_0}\right\|_{\mathrm{op}}^2 \vee \max_{\substack{I\in\mathcal{G}_k^{m_1}\\ J\in\mathcal{G}_q^{m_2}}}\left\|G_{IJ}\right\|_{\mathrm{op}}^2 \vee \max_{\substack{0\le i<k\\ 0\le j<q\\ (i,j)\neq(0,0)}} \frac{8}{\gamma\left(\frac{i}{k}+\frac{j}{q}\right)} \max_{\substack{|I\setminus I_0|=i\\ |J\setminus J_0|=j}}\left[\left\|G_{I\cap I_0,J\setminus J_0}^\top u_I\right\|_2^2 + \left\|G_{I\setminus I_0,J\cap J_0}v_J\right\|_2^2\right]. \tag{54}$$

Then, for every $G\in\mathbb{R}^{m_1\times m_2}$, the matrix $\Xi(G)$ defined in (53) belongs to the normal cone of $\Omega_{k,q}$ at A.

By choosing $\epsilon(G)$ as in Lemma 31, the upper bound (52) because $\Xi(G)\in N_{\Omega_{k,q}}(A)$. Using the decomposition $G=\Pi_{A,I_0,J_0}(G)+\Pi^{\perp}_{A,I_0,J_0}(G)$ we deduce

$$\begin{aligned}\mathfrak{S}(A,\Omega_{k,q}) \le \mathbb{E}\left\|G-\Xi(G)\right\|_{\mathrm{Fro}}^2 &= \mathbb{E}\left\|\epsilon(G)A-\Pi_{A,I_0,J_0}(G)\right\|_{\mathrm{Fro}}^2\\ &\le 2\mathbb{E}\left\|\epsilon(G)A\right\|_{\mathrm{Fro}}^2 + 2\mathbb{E}\left\|\Pi_{A,I_0,J_0}(G)\right\|_{\mathrm{Fro}}^2\\ &= 2\mathbb{E}\ \epsilon(G)^2 + 2(k+q-1), \end{aligned}\tag{55}$$

where (55) is due to $\|A\|_{\mathrm{Fro}}=1$ and the fact that $\|\Pi_{A,I_0,J_0}(G)\|_{\mathrm{Fro}}^2$ follows a chi-square distribution with $k+q-1$ degrees of freedom, since by Lemma 30 this is the dimension of $\mathrm{span}_{I_0,J_0}(A)$. In order to upper bound $\mathbb{E}\ \epsilon(G)^2$ we need the following two lemmata in addition to Lemma 28.

Lemma 32

$$\mathbb{E}\left\|G_{I_0,J_0}\right\|_{\mathrm{op}}^2 \le 4(k+q)+4\,. \tag{56}$$

Lemma 33

$$\begin{aligned}\mathbb{E}\max_{i,j}\frac{8}{\gamma\left(\frac{i}{k}+\frac{j}{q}\right)}\max_{\substack{|J\setminus J_0|=j\\ |I\setminus I_0|=i}}&\left[\|G_{I\cap I_0,J\setminus J_0}^\top u_I\|_2^2+\|G_{I\setminus I_0,J\cap J_0}v_J\|_2^2\right]\\ &\le \frac{48}{\gamma}(k\vee q)\log\left((m_1-k)\vee(m_2-q)\right)+\frac{64}{\gamma}(k\vee q)\,.\end{aligned}$$

Combining Lemmata 28, 56 and 33 with the definition of $\epsilon(G)$ in (54) we deduce

$$\begin{aligned}\mathbb{E}\ \epsilon(G)^2 &\le \frac{16}{\gamma^2}\left[4(k+q)+4\right]+16\left[\left(k\log\frac{m_1}{k}+q\log\frac{m_2}{q}\right)+2(k+q)\right]\\ &\qquad+\frac{48}{\gamma}(k\vee q)\log\left((m_1-k)\vee(m_2-q)\right)+\frac{64}{\gamma}(k\vee q)\\ &\le \left(\frac{64}{\gamma^2}+\frac{64}{\gamma}+32\right)(k+q+1)+16\left(k\log\frac{m_1}{k}+q\log\frac{m_2}{q}\right)\\ &\qquad+\frac{48}{\gamma}(k\vee q)\log\left(m_1\vee m_2\right)\\ &\le \frac{160}{\gamma^2}(k+q+1)+\frac{80}{\gamma}(k\vee q)\log\left(m_1\vee m_2\right)\,.\end{aligned}$$

Plugging this upper bound into (55) finally proves Proposition 23. ■

D.3 The scaling factor $\epsilon(G)$ ensures that $\Xi(G) \in N_{\Omega_{k,q}}(A)$ (proof of Lemma 31)

Proof [Lemma 31]
To simplify notations let us denote

$$\tilde{G} := \Pi^{\perp}_{A,I_0,J_0}(G)\,,$$

so that (53) becomes $\Xi(G) = \epsilon(G)A + \tilde{G}$. To prove that $\Xi(G)$ belongs to the normal cone of $\Omega_{k,q}$ at A, it is sufficient to prove that $\epsilon(G)^{-1}\Xi(G) = A + \epsilon(G)^{-1}\tilde{G}$ is a subgradient of $\Omega_{k,q}$ at A. By the characterization of the subgradient in (51), and since $\tilde{G} \in \mathrm{span}^{\perp}_{I_0,J_0}(A)$, this is equivalent to $\left\| A_{IJ} + \epsilon(G)^{-1}\,\tilde{G}_{IJ} \right\|_{\mathrm{op}} \leq 1$ for any $(I,J) \in \mathcal{G}_k^{m_1} \times \mathcal{G}_q^{m_2}$, which itself is equivalent to

$$\left\| A_{IJ} + \epsilon(G)^{-1}\,\Pi_{A,I,J}(\tilde{G}) \right\|_{\mathrm{op}} \leq 1 \qquad \text{and} \qquad \epsilon(G)^{-1}\,\left\| \mathcal{P}_A^{\perp}(\tilde{G}_{I,J}) \right\|_{\mathrm{op}} \leq 1\,. \tag{57}$$

First, the second inequality of (57) is satisfied since

$$\left\| \mathcal{P}_A^{\perp}(\tilde{G}_{I,J}) \right\|_{\mathrm{op}} \leq \left\| \tilde{G}_{I,J} \right\|_{\mathrm{op}} = \left\| \left[\Pi^{\perp}_{A,I_0,J_0}(G) \right]_{IJ} \right\|_{\mathrm{op}} \leq \|[G]_{IJ}\|_{\mathrm{op}} \leq \epsilon(G)\,.$$

There thus remains to prove the first inequality of (57). Note that the matrix $A_{IJ} + \epsilon(G)^{-1}\,\Pi_{A,I,J}(\tilde{G})$ has rank 2, so its Frobenius norm is larger than its operator norm by at most a factor of $\sqrt{2}$. Working with the Frobenius norm is more convenient, so knowing that

$$\left\| A_{IJ} + \epsilon(G)^{-1}\,\Pi_{A,I,J}(\tilde{G}) \right\|^2_{\mathrm{op}} \leq \left\| A_{IJ} + \epsilon(G)^{-1}\,\Pi_{A,I,J}(\tilde{G}) \right\|^2_{\mathrm{Fro}}\,,$$

we will establish an upper bound on the latter quantity which we denote by $\nu_{I,J}(G)$. Noting that $A_{IJ} = \|a_I\|_2\,\|b_J\|_2\,u_I v_J^{\top}$ and that

$$\Pi_{A,I,J}(\tilde{G}) = u_I u_I^{\top}\tilde{G}_{IJ} + \tilde{G}_{IJ}v_J v_J^{\top} - u_I u_I^{\top}\tilde{G}_{IJ}v_J\,v_J^{\top}\,,$$

we get

$$\begin{aligned}
\nu_{I,J}(G) &= \left\| \|a_I\|_2\,\|b_J\|_2\,u_I v_J^{\top} + \epsilon(G)^{-1}\left(u_I u_I^{\top}\tilde{G}_{IJ} + \tilde{G}_{IJ}v_J v_J^{\top} - u_I u_I^{\top}\tilde{G}_{IJ}v_J v_J^{\top} \right) \right\|^2_{\mathrm{Fro}} \\
&= \|a_I\|_2^2\,\|b_J\|_2^2 + \frac{2}{\epsilon(G)}\,\|a_I\|_2\,\|b_J\|_2\,u_I^{\top}\tilde{G}_{IJ}v_J \\
&\qquad + \frac{1}{\epsilon(G)^2}\left(u_I^{\top}\tilde{G}_{IJ}\tilde{G}_{IJ}^{\top}u_I + v_J^{\top}\tilde{G}_{IJ}^{\top}\tilde{G}_{IJ}v_J - 2(u_I^{\top}\tilde{G}_{IJ}v_J)^2 \right) \\
&\leq \|a_I\|_2^2\,\|b_J\|_2^2 + \frac{2}{\epsilon(G)}\,\|a_I\|_2\,\|b_J\|_2\,u_I^{\top}\tilde{G}_{IJ}v_J + \frac{1}{\epsilon(G)^2}\left(u_I^{\top}\tilde{G}_{IJ}\tilde{G}_{IJ}^{\top}u_I + v_J^{\top}\tilde{G}_{IJ}^{\top}\tilde{G}_{IJ}v_J \right)\,.
\end{aligned}$$

The following Lemma provides upper bounds on the different terms.

Lemma 34 *We have*

$$\begin{aligned}
u_I^{\top}\tilde{G}_{IJ}v_J &\leq \left\|a_{I_0\setminus I}\right\|_2\,\left\|b_{J_0\setminus J}\right\|_2\,\|G_{I_0J_0}\|_{\mathrm{op}}\,, \\
u_I^{\top}\tilde{G}_{IJ}\tilde{G}_{IJ}^{\top}u_I &\leq \left\|G^{\top}_{I\cap I_0,J\setminus J_0}u_I\right\|_2^2 + 2\,\left\|a_{I_0\setminus I}\right\|_2^2\,\|G_{I_0,J_0}\|^2_{\mathrm{op}}\,, \\
v_J^{\top}\tilde{G}_{IJ}^{\top}\tilde{G}_{IJ}v_J &\leq \left\|G_{I\setminus I_0,J\cap J_0}v_J\right\|_2^2 + 2\,\left\|b_{J_0\setminus J}\right\|_2^2\,\|G_{I_0,J_0}\|^2_{\mathrm{op}}\,.
\end{aligned}$$

This yields

$$\begin{aligned}\nu_{I,J}(G) \leq\ & \|a_I\|_2^2 \|b_J\|_2^2 + \frac{2}{\epsilon(G)} \|a_I\|_2 \|b_J\|_2 \left\|a_{I_0\setminus I}\right\|_2 \left\|b_{J_0\setminus J}\right\|_2 \|G_{I_0J_0}\|_{\mathrm{op}} \\ &+ \frac{1}{\epsilon(G)^2}\left(\left\|G^\top_{I\cap I_0, J\setminus J_0} u_I\right\|_2^2 + 2\left\|a_{I_0\setminus I}\right\|_2^2 \|G_{I_0,J_0}\|_{\mathrm{op}}^2\right) \\ &+ \frac{1}{\epsilon(G)^2}\left(\left\|G_{I\setminus I_0, J\cap J_0} v_J\right\|_2^2 + 2\left\|b_{J_0\setminus J}\right\|_2^2 \|G_{I_0,J_0}\|_{\mathrm{op}}^2\right) \\ \leq\ & \|a_I\|_2^2 \|b_J\|_2^2 + \frac{\gamma}{2}\|a_I\|_2 \|b_J\|_2 \left\|a_{I_0\setminus I}\right\|_2 \left\|b_{J_0\setminus J}\right\|_2 \\ &+ \frac{\gamma}{8}\left(\frac{i}{k}+\frac{j}{q}\right) + \frac{\gamma^2}{8}\left(\left\|a_{I_0\setminus I}\right\|_2^2 + \left\|b_{J_0\setminus J}\right\|_2^2\right),\end{aligned}$$

where we used the definition of $\epsilon(G)$ (54) to derive the last inequality.
Define $\alpha := \|a_{I_0\setminus I}\|^2 = 1 - \|a_I\|^2$ and $\beta := \|b_{J_0\setminus J}\|^2 = 1 - \|b_J\|^2$. With these notations and rearranging the terms, we can rewrite the above inequality as

$$\nu_{I,J}(G) \leq (1-\alpha)(1-\beta) + \frac{\gamma}{2}\sqrt{\alpha\beta(1-\alpha)(1-\beta)} + \frac{\gamma^2}{8}(\alpha+\beta) + \frac{\gamma}{8}\left(\frac{i}{k}+\frac{j}{q}\right).$$

Since $0 \leq \alpha, \beta \leq 1$ and using $\sqrt{\alpha\beta} \leq \frac{1}{2}(\alpha+\beta)$, we have

$$\alpha\beta \leq \frac{1}{2}(\alpha+\beta) \quad \text{and} \quad \sqrt{\alpha\beta(1-\alpha)(1-\beta)} \leq \frac{1}{2}(\alpha+\beta).$$

These inequalities yield

$$\nu_{I,J}(G) \leq 1 + (\alpha+\beta)\Big(-1+\frac{1}{2}+\frac{\gamma}{4}+\frac{\gamma^2}{8}\Big) + \frac{\gamma}{8}\left(\frac{i}{k}+\frac{j}{q}\right).$$

By definition of $\gamma = \min_{\substack{\iota\in I_0\\ \iota'\in J_0}} \left(k\, a_\iota^2, q\, b_{\iota'}^2\right)$, we have $\frac{i}{k} \leq \frac{\alpha}{\gamma}$ and $\frac{j}{q} \leq \frac{\beta}{\gamma}$. Moreover, given that $0 \leq \gamma \leq 1$, we have $\frac{4}{\gamma} - 2 - \gamma = \frac{1}{\gamma}(4 - 2\gamma - \gamma^2) \geq \frac{1}{\gamma}$, so that factorizing $\frac{\gamma}{8}$ in the previous expression, we obtain

$$\begin{aligned}\nu_{I,J}(G) &\leq 1 + \frac{\gamma}{8}\left[\Big(-\frac{4}{\gamma}+2+\gamma\Big)(\alpha+\beta) + \left(\frac{i}{k}+\frac{j}{q}\right)\right] \\ &\leq 1 + \frac{\gamma}{8}\left[-\frac{1}{\gamma}(\alpha+\beta) + \left(\frac{i}{k}+\frac{j}{q}\right)\right] \\ &\leq 1,\end{aligned}$$

which concludes the proof. ∎

D.4 Proof of Lemma 34

Let us first start with a few useful technical lemmas.

Lemma 35 *The matrix $\tilde{G}_{IJ} = [\Pi^\perp_{A,I_0,J_0}(G)]_{IJ}$ is of the form $\tilde{G}_{IJ} = \tilde{G}_1 + \tilde{G}_2$ with*

$$\tilde{G}_1 = G_{IJ} - G_{I\cap I_0, J\cap J_0} \quad \textit{and} \quad \tilde{G}_2 = (\mathrm{Id}_I - a_I a^\top)\, G_{I_0J_0}\, (\mathrm{Id}_J - bb_J^\top).$$

Proof [Lemma 35]

$$\begin{aligned} \Pi^{\perp}_{A,I_0,J_0}(G) &= G - \Pi_{A,I_0,J_0}(G) \\ &= G - a_{I_0}a_{I_0}^\top G_{I_0J_0} - G_{I_0J_0}b_{J_0}b_{J_0}^\top + a_{I_0}a_{I_0}^\top G_{I_0J_0}b_{J_0}b_{J_0}^\top \\ &= G - G_{I_0J_0} + (\mathrm{Id}_{I_0} - a_{I_0}a_{I_0}^\top)\, G_{I_0J_0}\, (\mathrm{Id}_{J_0} - b_{J_0}b_{J_0}^\top), \\ \text{so that} \quad [\Pi^{\perp}_{A,I_0,J_0}(G)]_{IJ} &= G_{IJ} - G_{I\cap I_0,J\cap J_0} + (\mathrm{Id}_I - a_I a^\top)\, G_{I_0J_0}\, (\mathrm{Id}_J - bb_J^\top). \end{aligned}$$

■

Lemma 36 *We have* $u_I^\top \tilde{G}_1 = u_I^\top G_{I\cap I_0,J\setminus J_0}$ *and* $\tilde{G}_1 v_J = G_{I\setminus I_0,J\cap J_0} v_J$.

Proof [Lemma 36]
Given that $\mathrm{supp}(u_I) \subset I_0$, we have

$$u_I^\top \tilde{G}_1 = u_I^\top (G_{IJ} - G_{I\cap I_0,J\cap J_0}) = u_I^\top (G_{I\cap I_0,J} - G_{I\cap I_0,J\cap J_0}) = u_I^\top G_{I\cap I_0,J\setminus J_0}\,,$$

which proves the first equality. The second one is proved similarly. ■

Lemma 37 $\quad \|\mathrm{Id} - b_J b^\top\|^2_{\mathrm{op}} \leq \dfrac{4}{3}$

Proof [Lemma 37]
The largest singular value is attained on the span of b_J and b_{J^c} both on the left and on the right. Given that $\|b\| = 1$, it is therefore also the largest eigenvalue of the matrix of the linear operator restricted to this span which is equal to

$$\begin{bmatrix} (1-x) & -\sqrt{(1-x)x} \\ 0 & 1 \end{bmatrix},$$

for $x = \|b_J\|^2$. Tedious but simple calculations show that the squared operator norm of this matrix is equal to $1 - x/2 + 1/2\sqrt{x(4-3x)}$, which takes its maximum value $4/3$ for $x = 1/3$. ■

Proof [Lemma 34]
Given that $\tilde{G}_{IJ} = \tilde{G}_1 + \tilde{G}_2$ and $u_I^\top \tilde{G}_1 = u_I \tilde{G}_{I\cap I_0,J\setminus J_0}$, we have $u_I^\top \tilde{G}_1 v_J = u_I^\top \tilde{G}_1 v_{J\cap J_0} = 0$, so that

$$\begin{aligned} u_I^\top \tilde{G}_{IJ} v_J &= u_I^\top \tilde{G}_2 v_J \\ &= u_I^\top (\mathrm{Id}_I - a_I a^\top)\, G_{I_0J_0}\, (\mathrm{Id}_J - bb_J^\top) v_J \\ &\leq \left\| u_I - \|a_I\|\, a \right\| \left\| G_{I_0J_0} \right\|_{\mathrm{op}} \left\| v_J - \|b_J\|\, b \right\| \\ &\leq \|a_{I_0\setminus I}\|\, \|b_{J_0\setminus J}\|\, \|G_{I_0J_0}\|_{\mathrm{op}}, \end{aligned}$$

because $\|u_I^\top(\mathrm{Id}_I - a_I a^\top)\|^2 = \left\|u_I - \|a_I\|\, a\right\|^2 = 1 - 2\|a_I\|^2 + \|a_I\|^2 = \|a_{I_0\setminus I}\|^2$, and symmetrically $\left\|v_J - \|b_J\|\, b\right\| = \|b_{J_0\setminus J}\|$. This shows the first inequality.
For the two next inequalities, note that

$$u_I^\top \tilde{G}_{IJ} \tilde{G}_{IJ}^\top u_I = \|\tilde{G}_{IJ}^\top u_I\|^2 = \|\tilde{G}_1^\top u_I\|^2 + \|\tilde{G}_2^\top u_I\|^2$$

because $\langle \tilde{G}_1^\top u_I, \tilde{G}_2^\top u_I \rangle = 0$ as a result of the fact that by lemma 36, $\tilde{G}_1^\top u_I$ and $\tilde{G}_2^\top u_I$ have disjoint supports.
Now $\|\tilde{G}_1^\top u_I\|^2 = \|G^\top_{I\cap I_0, J\setminus J_0} u_I\|_2^2$ and $\|\tilde{G}_2^\top u_I\| \leq 2\,\|a_{I_0\setminus I}\|^2\,\|G_{I_0,J_0}\|_{\mathrm{op}}^2$, because $\|\mathrm{Id} - b_J b^\top\|_{\mathrm{op}}^2 \leq 2$ (see Lemma 37 for a proof). This shows the second inequality and the third follows by symmetry. ■

D.5 Upper bounds for $\epsilon(G)^2$ (Proofs of Lemmata 32 and 33)

Proof [Lemma 32]

Using (45) and the fact that $\left(\sqrt{k}+\sqrt{q}+s\right)^2 \leq 2\left((\sqrt{k}+\sqrt{q})^2+s^2\right)$ gives

$$\mathbb{P}\left[\|G_{I_0,J_0}\|_{\mathrm{op}}^2 > 2\left((\sqrt{k}+\sqrt{q})^2+s^2\right)\right] \leq \exp(-s^2/2)\,.$$

Setting $t = 2s^2$ yields

$$\mathbb{P}\left[\|G_{I_0,J_0}\|_{\mathrm{op}}^2 > 4(k+q)+t\right] \leq \exp(-t/4)\,.$$

It follows that

$$\begin{aligned}
\mathbb{E}\ \|G_{I_0,J_0}\|_{\mathrm{op}}^2 &= \int_0^\infty \mathbb{P}(\|G_{I_0,J_0}\|_{\mathrm{op}}^2 \geq t')dt' \\
&= \int_0^{4(k+q)} dt' + \int_{4(k+q)}^\infty \mathbb{P}(\|G_{I_0,J_0}\|_{\mathrm{op}}^2 \geq t')dt' \\
&\leq 4(k+q) + \int_0^\infty \exp(-t/4)dt \\
&= 4(k+q)+4\,.
\end{aligned}$$

■

Proof [Lemma 33]
As the sets $I\cap I_0 \times J\setminus J_0$ and $I\setminus I_0 \times J\cap J_0$ are disjoint, and u_I, v_J of unit length, the random variable

$$M_{I,J} = \left\|G^\top_{I\cap I_0, J\setminus J_0} u_I\right\|_2^2 + \left\|G_{I\setminus I_0, J\cap J_0} v_J\right\|_2^2$$

follows a chi-square distribution with $i+j$ degrees of freedom, where $i = |I\setminus I_0|$ and $j = |J\setminus J_0|$. Using Chernoff's inequality and the form of the chi-square moment generating function, we have that for any fixed real number α and fixed index sets I and J, for all $t \in (0, 1/2)$,

$$\mathbb{P}\Big[M_{I,J} > \alpha\Big] = \mathbb{P}\Big[e^{tM_{I,J}} > e^{t\alpha}\Big] \leq e^{-t\alpha}\ \mathbb{E}\ e^{tM_{I,J}} = e^{-t\alpha}(1-2t)^{-\frac{i+j}{2}}\,.$$

Taking the maximum over index sets I and J with the same intersection sizes with I_0 and J_0 respectively, and using a union bound on the independent choices of I and J, yields

$$\begin{aligned}
\mathbb{P}\left[\max_{\substack{|I\setminus I_0|=i\\ |J\setminus J_0|=j}} M_{I,J} > \alpha\right] &\leq \binom{m_1-k}{i}\binom{m_2-q}{j}\exp\left\{-t\alpha - \frac{i+j}{2}\log(1-2t)\right\} \\
&\leq \exp\left\{-t\alpha - \frac{i+j}{2}\log(1-2t) + i\,\log(m_1-k) + j\,\log(m_2-q)\right\}.
\end{aligned}$$

Taking $\alpha = \lambda(i+j)$, we have for any $t < 1/2$ (assuming w.l.o.g. $m_1 - k \geq m_2 - q$)

$$\mathbb{P}\left[\max_{\substack{|I\setminus I_0|=i\\|J\setminus J_0|=j}} M_{I,J} > \lambda(i+j)\right] \leq \exp\left\{-t\lambda(i+j) - \frac{i+j}{2}\log(1-2t) + i\,\log(m_1-k) + j\,\log(m_2-q)\right\}$$
$$\leq \exp\left\{(i+j)\left(-t\lambda - \frac{1}{2}\log(1-2t) + \log(m_1-k)\right)\right\}.$$

Let us introduce $\mathcal{M}_{i,j} = \frac{1}{i+j}\max_{\substack{|I\setminus I_0|=i\\|J\setminus J_0|=j}} M_{I,J}$, and take $t = \frac{1}{2}\left(1 - \frac{1}{m_1-k}\right) < \frac{1}{2}$. Then

$$\mathbb{P}\left[\max_{\substack{0\leq i<k\\0\leq j<q\\(i,j)\neq(0,0)}} \mathcal{M}_{i,j} > \lambda\right] \leq \sum_{\substack{0\leq i<k\\0\leq j<q\\(i,j)\neq(0,0)}} \exp\left\{(i+j)\left(-\frac{1}{2}\left(1-\frac{1}{m_1-k}\right)\lambda + \frac{3}{2}\log(m_1-k)\right)\right\}$$
$$= \sum_{i=0}^{k-1}\beta^i\sum_{j=0}^{q-1}\beta^j - 1 = \frac{1-\beta^k}{1-\beta}\frac{1-\beta^q}{1-\beta} - 1 \leq 2\beta\,,$$

where

$$\beta = \exp\left\{-\frac{1}{2}\left(1-\frac{1}{m_1-k}\right)\lambda + \frac{3}{2}\log(m_1-k)\right\}.$$

As a consequence, we have

$$\begin{aligned}
\mathbb{E}[\max_{i,j}\mathcal{M}_{i,j}] &= \int_0^\infty \mathbb{P}[\max_{i,j}\mathcal{M}_{i,j} > \lambda]d\lambda\\
&\leq \int_0^{\frac{3(m_1-k)}{m_1-k-1}\log k} d\lambda + 2\int_{\frac{3(m_1-k)}{m_1-k-1}\log(m_1-k)}^\infty \exp\left\{\frac{3}{2}\log(m_1-k) - \frac{1}{2}\left(1-\frac{1}{m_1-k}\right)\lambda\right\}d\lambda\\
&\leq \frac{3(m_1-k)}{m_1-k-1}\log k + 4\frac{m_1-k}{m_1-k-1}\\
&\leq 6\log(m_1-k) + 8\,.
\end{aligned}$$

It follows that

$$\mathbb{E}\max_{\substack{0\leq i<k\\0\leq j<q\\(i,j)\neq(0,0)}} \frac{8}{\gamma\left(\frac{i}{k}+\frac{j}{q}\right)}\max_{\substack{|J\setminus J_0|=j\\|I\setminus I_0|=i}} \|G^\top_{I\cap I_0, J\setminus J_0}u_I\|_2^2 + \|G_{I\setminus I_0, J\cap J_0}v_J\|_2^2$$
$$\leq \frac{48}{\gamma}(k\vee q)\log\left((m_1-k)\vee(m_2-q)\right) + \frac{64}{\gamma}(k\vee q)\,. \tag{58}$$

■

E Lower bound on the statistical dimension of Γ_μ (Proof of Proposition 24)

Let us start with a technical lemma:

Lemma 38 *Let $ab^\top \in \mathcal{A}_{k,q}$, $\mathcal{X} : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^n$ a linear map from the standard Gaussian ensemble and $y = \mathcal{X}(ab^\top)$. If $n \leq \frac{1}{9} m_1 m_2$ and further*

$$n \leq n_0 := \zeta(a,b)\,\frac{1}{6^4}\big((kq) \wedge (m_1 + m_2 - 1)\big) - 2, \quad \textit{with} \quad \zeta(a,b) = 1 - \Big(1 - \frac{\|a\|_1^2}{k}\Big)\Big(1 - \frac{\|b\|_1^2}{q}\Big),$$

then, with probability $1 - c_1 \exp(-c_2 n_0)$, solving formulation (28) with the norm Γ_μ fails to recover $ab^\top$ simultaneously for any values of $\mu \in [0,1]$, where c_1 and c_2 are universal constants.

Proof [Lemma 38]
The proof consists in applying theorem 3.2 in Oymak et al. (2012) for the combination of the ℓ_1-norm with the trace norm. We adapt slightly the notations of that paper to reflect the fact that we are working with matrices. Since we consider conic combinations of the ℓ_1 and trace norms, the number of norms is therefore $\tau = 2$. To apply the theorem we need to specify $\kappa, \theta, d_{\min}, \gamma$ and $\mathcal{C}^\circ$ in the notations of that paper.
For each decomposable norm ν_j for $j \in \{1,2\}$, with ν_1 the ℓ_1-norm and and ν_2 the trace norm, given a point $ab^\top$ (which corresponds to the point $\mathbf{x}_0$ in Oymak et al., 2012), the authors define

- T_j the supporting subspaces and E_j ($\mathbf{e}_j$ in the paper), the orthogonal projection of any subgradient of the norm in $ab^\top$ (Definition 2.1),
- L_j the Lipschitz constant of ν_j with respect to the Euclidean norm (Definition 2.2),
- $\kappa_j = \dfrac{\|E_j\|_{\text{Fro}}^2}{L_j^2} \dfrac{m_1 m_2}{\dim(T_j)}$ (Definition 2.2).

Let $ab^\top \in \mathcal{A}_{k,q}$ with support $I_0 \times J_0$ and $s_a = \text{sign}(a)$, $s_b = \text{sign}(b)$. Denoting e_{ij} the element of the canonical basis of $\mathbb{R}^{m_1 \times m_2}$, we have

- $T_1 = \text{span}(\{e_{ij}\}_{(i,j) \in I_0 \times J_0})$ so that $\dim(T_1) = kq$,
- $T_2 = \{av^\top + ub^\top \mid u \in \mathbb{R}^{m_1}, v \in \mathbb{R}^{m_2}\}$ so that $\dim(T_2) = m_1 + m_2 - 1$.

By definition $d_{\min} = \dim(T_1) \wedge \dim(T_2)$. We have

$$E_1 = s_a s_b^\top, \;\; \|E_1\|_{\text{Fro}}^2 = kq, \;\; E_2 = ab^\top, \;\; \|E_2\|_{\text{Fro}}^2 = 1, \;\; L_1 = \sqrt{kq}, \;\; L_2 = \sqrt{m_1 \wedge m_2},$$

$$\text{and thus} \quad \kappa_1 = \frac{m_1 m_2}{kq}, \;\; \kappa_2 = \frac{m_1 m_2}{(m_1 \wedge m_2)(m_1 + m_2 - 1)}, \;\; \text{so that } \kappa = \kappa_1 \wedge \kappa_2 \geq \frac{1}{2}.$$

We then have θ defined as $\theta = \theta_1 \wedge \theta_2$ with $\theta_j = \|E_{\cap,j}\|_2 / \|E_j\|_2$ where $E_{\cap,j}$ is the projection of E_j on $T_1 \cap T_2$. But $E_2 \in T_1$ so that $\theta_2 = 1$. The situation is less simple for E_1. Indeed, $E_{\cap,1} = \|a\|_1 a s_b^\top + \|b\|_1 s_a b^\top - ab^\top \|a\|_1 \|b\|_1$. Some calculations lead to

$$\theta_1^2 = \frac{\|a\|_1^2}{k} + \frac{\|b\|_1^2}{q} - \frac{\|a\|_1^2}{k}\frac{\|b\|_1^2}{q},$$

hence the definition of $\zeta(a,b) = \theta^2 = \theta_1^2 \wedge \theta_2^2$. Theorem 3.2 in Oymak et al. (2012) offers the possibility of constraining the estimator to lie in a cone $\mathcal{C}$. In our case, $\mathcal{C} = \mathbb{R}^{m_1 \times m_2}$, given the definition of γ we therefore have $\gamma \leq 2$. The result follows from applying the theorem with $\theta^2 = \zeta(a,b)$ and using $\frac{\kappa}{81\gamma^2\tau} \geq \frac{1/2}{3^4.2^2.2} = \frac{1}{6^4}$. ■

Proof [Proposition 24]
Take M such that when $m_1, m_2, k, q, m1/k, m_2/q \geq M$ then n_0 is large enough to ensure $1 - c_1 \exp(-c_2 n_0) > 4\exp(-32/17)$. Then, according to Lemma 38, solving (28) with the norm Γ_μ fails to recover $A = ab^\top$ with probability at least $4\exp(-32/17)$. On the other hand, Amelunxen et al. (2013, Theorem 7.1) shows that, when $n \geq \mathfrak{S}(A, \Gamma_\mu) + \lambda$, for any $\lambda \geq 0$, then solving (28) with the norm Γ_μ correctly recovers A with probability at least

$$4\exp\left(\frac{-\lambda^2/8}{\omega^2(A,\Gamma_\mu)+\lambda}\right), \tag{59}$$

where $\omega^2(A, \Gamma_\mu) = \mathfrak{S}(A, \Gamma_\mu) \wedge (m_1 m_2 - \mathfrak{S}(A, \Gamma_\mu))$. Take $\lambda = 16\omega(A, \Gamma_\mu)$, then using the fact that $\omega(A, \Gamma_\mu) \geq 1$ we get that the probability (59) is smaller than $4\exp(-32/17)$. This implies that

$$n_0 \leq \mathfrak{S}(A,\Gamma_\mu) + \lambda \leq \mathfrak{S}(A,\Gamma_\mu) + 16\sqrt{\mathfrak{S}(A,\Gamma_\mu)} \leq 17\mathfrak{S}(A,\Gamma_\mu)\,.$$

■

F Bounds on the statistical dimension in the vector case (proofs of results of Section 4.2.4)

F.1 Lower bound on the statistical dimension of κ_k (Proof of Proposition 25)

Let us start with two technical lemmata.

Lemma 39 *Let $X_{(k)}$ denote the kth order statistics of an i.i.d. sample $X_1, \ldots, X_n$ whose common distribution has a cdf F. Assume that F^{-1} is a convex function[7] from $[0,1]$ to $\overline{\mathbb{R}}$. Then*

$$\mathbb{E}[X_{(k)}] \geq F^{-1}\Big(\frac{k}{n+1}\Big)\,.$$

Proof [Lemma 39]
Let f denote the pdf of X. We have

$$\begin{aligned}\mathbb{E}[X_{(k)}] &= \frac{n!}{(k-1)!(n-k)!}\int_{-\infty}^{\infty} u\,F(u)^{k-1}\big(1-F(u)\big)^{n-r} f(u)\,du \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\,\Gamma(n-k+1)}\int_0^1 F^{-1}(v)\,v^{k-1}(1-v)^{n-r}dv = \mathbb{E}[F^{-1}(V)]\,,\end{aligned}$$

with $V \sim \mathrm{Beta}(k, n-k+1)$. Assuming that F^{-1} is a convex function, we have by Jensen's inequality

$$\mathbb{E}[X_{(k)}] = \mathbb{E}[F^{-1}(V)] \geq F^{-1}(\mathbb{E}[V]) = F^{-1}\Big(\frac{k}{n+1}\Big)\,.$$

■

[7] Note that this implies that the essential support of the random variable is bounded below.

Lemma 40 *Let $G \in \mathbb{R}^n$ be an standard normal vector, then we have*

$$\mathbb{E}[\kappa_k^*(G)] \geq \sqrt{\frac{2}{\pi}}\sqrt{k \log\Big(\frac{n+1}{k+1}\Big)}\,.$$

Proof [Lemma 40]
Denote by F the cdf of the absolute value of a standard normal variable. Then,

$$F(x) = \Phi(x) - \Phi(-x) = \mathrm{erf}\Big(\frac{x}{\sqrt{2}}\Big)\,,$$

where Φ is the cdf of a standard Gaussian and erf denotes the error function. We use the following inequality due to Chu (1954):

$$\sqrt{1 - e^{-x^2}} \leq \mathrm{erf}(x) \leq \sqrt{1 - e^{-\frac{\pi}{4}x^2}}\,,$$

to deduce that

$$F^{-1}(y) \geq \sqrt{-\frac{2}{\pi}\log(1-y^2)}\,.$$

By definition, we have $\mathbb{E}[\kappa_k^*(G)] = \frac{1}{\sqrt{k}}\mathbb{E}[X_{(n)} + \ldots + X_{(n-k+1)}]$ where $X_i = |G_i|$ and G is a vector of independent standard normal variables. It can easily be checked that F^{-1} is a convex function. This implies, using Lemma 39, that

$$\begin{aligned}
\mathbb{E}[\kappa_k^*(G)] &\geq \frac{1}{\sqrt{k}}\sum_{j=1}^{k} F^{-1}\Big(1 - \frac{j}{n+1}\Big) \\
&\geq \sqrt{k}\,F^{-1}\Big(\frac{1}{k}\sum_{j=1}^{k}\Big(1 - \frac{j}{n+1}\Big)\Big) \qquad \text{(again by Jensen's inequality)} \\
&= \sqrt{k}\,F^{-1}\Big(1 - \frac{k+1}{2(n+1)}\Big) \\
&\geq \sqrt{k}\sqrt{-\frac{2}{\pi}\log\Big(\frac{k+1}{(n+1)} - \Big(\frac{k+1}{2(n+1)}\Big)^2\Big)} \\
&\geq \sqrt{\frac{2}{\pi}}\sqrt{k\,\log\Big(\frac{n+1}{k+1}\Big)}\,.
\end{aligned}$$

■

Proof [Proposition 25]
We will denote the squared Gaussian width of the tangent cone intersected with a Euclidean unit ball by

$$w(T_{\kappa_k}(a) \cap \mathbb{S}^{m-1}) = \mathbb{E}\Big[\max_{t \in T_{\kappa_k}(a)\cap\mathbb{S}^{m-1}} \langle t, G\rangle\Big]\,,$$

where $G \in \mathbb{R}^m$ denotes a standard Gaussian vector. We have $w(T_{\kappa_k}(a)\cap\mathbb{S}^{m-1})^2 \leq \mathfrak{S}(a,\kappa_k)$(Chandrasekaran et al., 2012, Proposition 3.6). We thus seek a lower bound of $w(T_{\kappa_k}(a) \cap \mathbb{S}^{m-1})$. Since the tangent cone is polar to the normal cone, we have that

$$T_{\kappa_k}(a) = \{t \in \mathbb{R}^m \mid \langle s, t\rangle \leq 0,\ \forall s \in \partial\kappa_k(a)\}\ .$$

Given a random Gaussian vector G, denote I_0 the support of a and I_G the indices of the k largest coefficients of G in absolute value outside of I_0. Denote by $\tilde{s}_G = \mathrm{sign}(G_{I_G})$, *i.e.*, the vector whose entries are zero outside of I_G and equal to the sign of the corresponding coefficient of G otherwise. Define $t_G = \frac{1}{\sqrt{2k}}(\tilde{s}_G - a)$. By construction $t_G \in \mathbb{S}^{m-1}$. Let now consider $s \in \partial\kappa_k(a)$, we have

$$\sqrt{2k}\,\langle s, t_G\rangle \;=\; -\langle s, a\rangle + \langle s, \tilde{s}_G\rangle \leq -1 + \kappa_k(\tilde{s}_G)\,\kappa_k^*(s) \leq -1 + 1 = 0,$$

so that $t_G \in T_{\kappa_k}(a)$. Therefore $w(T_{\kappa_k}(a) \cap \mathbb{S}^{m-1}) \geq \mathbb{E}[\langle t_G, G\rangle] = \frac{1}{2\sqrt{k}}\mathbb{E}[\langle \tilde{s}_G, G\rangle] = \frac{1}{2}\mathbb{E}[\kappa_k^*(G)]$, whence the result using Lemma 40 and $w(T_{\kappa_k}(a) \cap \mathbb{S}^{m-1})^2 \leq \mathfrak{S}(a, \kappa_k)$. ■

F.2 Upper bound on the statistical dimension of θ_k (Proof of Proposition 26)

Proof [Proposition 26]
Without loss of generality, let us assume that $w \in \mathbb{R}^p$ is a fixed vector having nonincreasing – in absolute value – coordinates, the first s of which are assumed to be nonzero. We compute the subdifferential of $\theta_k(w)$ directly by using (14). Remember that one characterization of the subdifferential is

$$\partial\theta_k(w) = \{\alpha \in \mathbb{R}^p \;:\; \theta_k^*(\alpha) \leq 1\,,\; \alpha^\top w = \theta_k(w)\}\;.$$

Letting $r \in \{0, \cdots, k-1\}$ being the unique integer such that $|w_{k-r-1}| > \frac{1}{r+1}\sum_{i=k-r}^{p}|w_i| \geq |w_{k-r}|$, let us partition the set of entries $\{1, \cdots, p\}$ into $I_2 = \{1, \cdots, k-r-1\}$, $I_1 = \{k-r, \cdots, s\}$ and $I_0 = \{s+1, \cdots, p\}$ (where each set may be empty). Then we can rewrite the expression of the k-support norm (14) as

$$\theta_k(w)^2 = \|w_{I_2}\|_2^2 + \frac{1}{r+1}\|w_{I_1}\|_1^2\,.$$

Then necessarily each element $\alpha \in \partial\theta_k(w)$ must satisfy

$$\begin{cases} \alpha_i = \frac{w_i}{\theta_k(w)} & \text{for } i \in I_2\,, \\ \alpha_i = \frac{\|w_{I_1}\|_1 \,\mathrm{sign}(w_i)}{(r+1)\theta_k(w)} & \text{for } i \in I_1\,. \end{cases}$$

As for $i \in I_0$, the coefficients α_i do not impact $\alpha^\top w$ so they should also not impact $\theta_k^*(\alpha)$. If $s < k$ this implies $\alpha_i = 0$, and if $s \geq k$ this means $|\alpha_i| \leq |\alpha_k|$, and in that case $k \in I_1$. With the convention $\|w_{I_1}\|_1 = 0$ when $I_1 = \emptyset$, we finally get the following expression for the subdifferential:

$$\partial\theta_k(w) = \frac{1}{\theta_k(w)}\left\{w_{I_2} + \frac{1}{r+1}\,\|w_{I_1}\|_1\,(\mathrm{sgn}(w_{I_1}) + h_{I_0}) \;:\; \|h\|_\infty \leq 1\right\}. \tag{60}$$

In the case $s < k$, we have $s = k - r + 1$, $I_2 = [1, s]$, $I_1 = \emptyset$ and $I_0 = [s+1, p]$. In that case $\theta_k(w) = \|w\|_2$ and $\partial\theta_k(w) = w/\|w\|_2$, showing that θ_k is differentiable at w, meaning θ_k is useless to recover w.
Let us therefore only consider the case $s \geq k$, in which case $I_1 \neq \emptyset$ and $\|w_{I_1}\|_1 > 0$. In order to compute the statistical dimension of θ_k at w, we use the characterization (50)

$$\mathfrak{S}(w, \theta_k) = \mathbb{E}\left[\mathrm{dist}\,(g, N_{\theta_k}(A))^2\right]\,,$$

where g is a p-dimensional random vector with i.i.d. normal entries and $N_{\theta_k}(A)$ is the conic hull of $\partial\theta_k(w)$. We then get:

$$\begin{aligned}
\mathfrak{S}(w,\theta_k) &= \mathbb{E}\left[\inf_{t>0\ \&\ u\in t\partial\theta_k(w)} \|u-g\|_2^2\right] \\
&\le \inf_{t>0}\ \mathbb{E}\left[\inf_{u\in t\partial\theta_k(w)} \|u-g\|_2^2\right] \\
&\le \inf_{t>0}\ \mathbb{E}\inf_{h\in\mathbb{R}^p,\|h\|_\infty\le 1}\left\{\left\|g_{I_2} - t\frac{(r+1)}{\|w_{I_1}\|_1}w_{I_2}\right\|_2^2 + \|g_{I_1} - t\,\mathrm{sgn}(w_{I_1})\|_2^2\right. \\
&\qquad\qquad \left. + \|g_{I_0} - th_{I_0}\|_2^2\right\} \\
&\le \inf_{t>0}\left\{|I_2| + \frac{(r+1)^2\|w_{I_2}\|_2^2}{\|w_{I_1}\|_1^2}t^2 + |I_1|(1+t^2) + |I_0|\frac{2}{\sqrt{2\pi}}\frac{1}{t}\exp\left(-\frac{t^2}{2}\right)\right\} \qquad (61) \\
&= \inf_{t>0}\ \left\{s + t^2\left\{\frac{(r+1)^2\|w_{I_2}\|_2^2}{\|w_{I_1}\|_1^2} + |I_1|\right\} + (p-s)\frac{2}{\sqrt{2\pi}}\frac{1}{t}\exp\left(-\frac{t^2}{2}\right)\right\} \\
&\le \frac{5}{4}s + 2\left\{\frac{(r+1)^2\|w_{I_2}\|_2^2}{\|w_{I_1}\|_1^2} + |I_1|\right\}\log\frac{p}{s}\,, \qquad (62)
\end{aligned}$$

where following Chandrasekaran et al. (2012, Annex C), for (61) we used the fact that for a standard normal random variable $G\sim\mathcal{N}(0,1)$

$$\mathbb{E}_G\inf_{|\eta|\le 1}(G-t\eta)^2 \le \frac{2}{\sqrt{2\pi}}\frac{1}{t}e^{-\frac{t^2}{2}}\,,$$

while (62) is obtained by taking $b=\sqrt{2\log(p/s)}$ and using $\frac{s(1-s/p)}{\sqrt{\pi\log(p/s)}}\le\frac{1}{4}$.
For the lasso case ($k=1$), we have $r=0$, $I_2=\emptyset$ and $I_1=[1,s]$. Plugging this into (62) we recover the standard bound (Chandrasekaran et al., 2012):

$$\mathfrak{S}(w,\theta_k) \le \frac{5}{4}s + 2s\log\frac{p}{s}\,. \qquad (63)$$

In the general case $1\le k\le s$ remember that, by definition of r,

$$|w_{k-r-1}| > \frac{\|w_{I_1}\|_1}{r+1} \ge |w_{k-r}|\,,$$

and therefore

$$|I_2| \le \frac{\|w_{I_2}\|_2^2}{|w_{k-r-1}|^2} \le \frac{(r+1)^2\|w_{I_2}\|_2^2}{\|w_{I_1}\|_1^2} \le \frac{\|w_{I_2}\|_2^2}{|w_{k-r}|^2}\,. \qquad (64)$$

Plugging the left-hand inequality of (64) into (62) and remembering that $|I_2|+|I_1|=s$ shows that the bound (62) obtained for θ_k, for any $1\le k\le s$, is never better than the bound (63) obtained for the lasso case $k=1$. In the case $s=k$, the right-hand inequality of (64) applied to an atom $w\in\mathcal{A}_p^k$ with atom strength $\gamma=k|w_k|^2$ and unit ℓ_2 norm leads to

$$\frac{(r+1)^2\|w_{I_2}\|_2^2}{\|w_{I_1}\|_1^2} + |I_1| \le \frac{\|w_{I_2}\|_2^2}{|w_{k-r}|^2} + |I_1| \le \frac{\|w_{I_2}\|_2^2}{|w_k|^2} + \frac{\|w_{I_1}\|_2^2}{|w_k|^2} = \frac{1}{|w_k|^2} = \frac{k}{\gamma}\,,$$

from which we deduce by (62) the upper bound on the statistical dimension

$$\forall w \in \mathcal{A}_p^k\,, \quad \mathfrak{S}(w, \theta_k) \leq \frac{5}{4}k + \frac{2k}{\gamma}\log\frac{p}{k}\,.$$

■