# Relaxations convexes pour l'estimation de matrices à facteurs parcimonieux

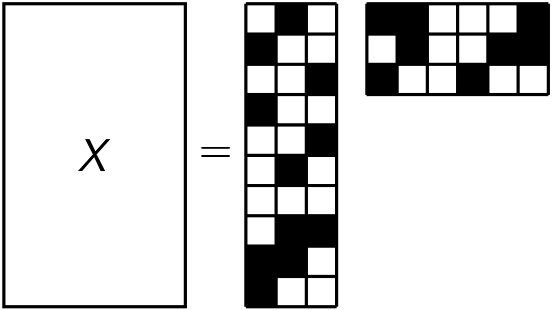
Guillaume Obozinski

Ecole des Ponts - ParisTech

Collaboration avec Emile Richard et Jean-Philippe Vert

Journées MAS, Toulouse, Août 2014

# Estimation low rank matrices with sparse factors



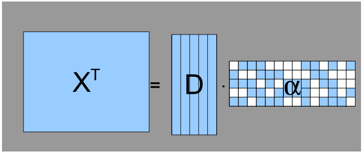
$$X = \sum_{i=1}^{r} u_i v_i^\top$$

- factors not orthogonal a priori
- $\neq$ from assuming the SVD of $X$ is sparse

# Dictionary Learning

$$\min_{\substack{A\in\mathbb{R}^{k\times n}\\ D\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|x_i - D\alpha_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\alpha_i\|_1 \quad \text{s.t.} \quad \forall j,\ \|d_j\|_2 \leq 1.$$

## Dictionary Learning

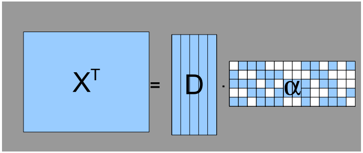


- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)

# Dictionary Learning /Sparse PCA

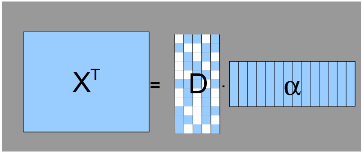
$$\min_{\substack{A\in\mathbb{R}^{k\times n}\\ D\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|x_i - D\alpha_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\alpha_i\|_1 \quad \text{s.t.} \quad \forall j,\ \|d_j\|_2 \leq 1.$$

## Dictionary Learning



- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)

## Sparse PCA



- e.g. microarray data
- sparse dictionary
- (Witten et al., 2009; Bach et al., 2008)

**Sparsity of the loadings vs sparsity of the dictionary elements**

# Applications

## Low rank factorization with “community structure”

Modeling clusters or community structure in social networks or recommendation systems (Richard et al., 2012).

## Subspace clustering (Wang et al., 2013)

Up to an unknown permutation, $X^\top = \begin{bmatrix} X_1^\top & \dots & X_K^\top \end{bmatrix}$ with $X_k$ low rank, so that there exists a low rank matrix $Z_k$ such that $X_k = Z_k X_k$. Finally,

$$X = ZX \qquad \text{with} \qquad Z = \mathrm{BkDiag}(Z_1, \dots, Z_K).$$

## Sparse PCA from $\hat{\Sigma}_n$

## Sparse bilinear regression

$$y = x^\top M x' + \varepsilon$$

# Existing approaches

## Bi-convex formulations

$$\min_{U,V} \mathcal{L}(UV^\top) + \lambda(\|U\|_1 + \|V\|_1),$$

with $U \in \mathbb{R}^{n\times r},\ V \in \mathbb{R}^{p\times r}$.

# Existing approaches

## Bi-convex formulations

$$\min_{U,V} \mathcal{L}(UV^\top) + \lambda(\|U\|_1 + \|V\|_1),$$

with $U \in \mathbb{R}^{n\times r},\ V \in \mathbb{R}^{p\times r}$.

## Convex formulation for sparse **and** low rank

$$\min_{Z} \mathcal{L}(Z) + \lambda\|Z\|_1 + \mu\|Z\|_*$$

- Doan and Vavasis (2013); Richard et al. (2012)
- factors not necessarily sparse as $r$ increases.

# Rank one case with square loss

$$\min_{u,v,\sigma} \|X - \sigma u v^\top\|_2 \qquad \text{s.t.} \qquad \begin{aligned} &\sigma \in \mathbb{R}_+, \\ &u \in \mathcal{U} \subset \mathbb{R}^n, \quad v \in \mathcal{V} \subset \mathbb{R}^p, \\ &\|u\|_2 = \|v\|_2 = 1. \end{aligned}$$

is equivalent to solving

$$\max_{u,v} \quad u^\top X v \qquad \text{s.t.} \qquad \begin{aligned} &u \in \mathcal{U} \subset \mathbb{R}^n, \quad v \in \mathcal{V} \subset \mathbb{R}^p, \\ &\|u\|_2 = \|v\|_2 = 1. \end{aligned}$$

- Corresponds to sparse PCA when $u = v$.

# Convex relaxations for sparse PCA

## Approaches differ according to view

- analysis view → build sequences of rank 1 approximations,
- synthesis view → find a set of common factors simultaneously

# Convex relaxations for sparse PCA

## Approaches differ according to view

- analysis view → build sequences of rank 1 approximations,
- synthesis view → find a set of common factors simultaneously

## Analysis SPCA focusses on solving rank-1 sparse PCA

- convex formulations: d'Aspremont et al. (2007, 2008); Amini and Wainwright (2009)
- modified power methods: Journée et al. (2010); Luss and Teboulle (2013); Yuan and Zhang (2013)

# Convex relaxations for sparse PCA

## Approaches differ according to view

- analysis view → build sequences of rank 1 approximations,
- synthesis view → find a set of common factors simultaneously

## Analysis SPCA focusses on solving rank-1 sparse PCA

- convex formulations: d'Aspremont et al. (2007, 2008); Amini and Wainwright (2009)
- modified power methods: Journée et al. (2010); Luss and Teboulle (2013); Yuan and Zhang (2013)

## Synthesis SPCA focusses on finding several complementary sparse factors

Essentially based on nuclear norms (Jameson, 1987; Bach et al., 2008; Bach, 2013).

# A new formulation for sparse matrix factorization and a new matrix norm

# A new formulation for sparse matrix factorization

Assumptions:

$$X = \sum_{i=1}^{r} a_i b_i^\top$$

- All left factors $a_i$ have support of size $k$.
- All right factors $b_i$ have support of size $q$.

# A new formulation for sparse matrix factorization

## Assumptions:

$$X = \sum_{i=1}^{r} a_i b_i^\top$$

- All left factors $a_i$ have support of size $k$.
- All right factors $b_i$ have support of size $q$.

## Goals:

Propose a convex formulation for sparse matrix factorization that

- is able to handle multiple sparse factors
- permits to identify the sparse factors themselves
- leads to better statistical performance than $\ell_1$/trace norm.

# A new formulation for sparse matrix factorization

## Assumptions:

$$X = \sum_{i=1}^{r} a_i b_i^\top$$

- All left factors $a_i$ have support of size $k$.
- All right factors $b_i$ have support of size $q$.

## Goals:

Propose a convex formulation for sparse matrix factorization that

- is able to handle multiple sparse factors
- permits to identify the sparse factors themselves
- leads to better statistical performance than $\ell_1$/trace norm.

Propose algorithms based on this formulation.

# $(k, q)$-sparse counterpart of the rank

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

# $(k, q)$-sparse counterpart of the rank

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

Given a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$,

# $(k, q)$-sparse counterpart of the rank

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

Given a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, consider

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

# $(k, q)$-sparse counterpart of the rank

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

Given a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, consider

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

Define

the $(k, q)$-rank of $Z$ as the optimal value $r := \|c^\star\|_0$

# $(k, q)$-sparse counterpart of the rank

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

Given a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, consider

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

Define

the $(k, q)$-rank of $Z$ as the optimal value $r := \|c^\star\|_0$

a $(k, q)$-decomposition of $Z$ any optimal solution $(a_i^\star, b_i^\star, c_i^\star)_{1 \leq i \leq r}$

For a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, we have

| | $(1,1)$-rank | $(k,q)$-rank | $(m_1,m_2)$-rank |
|---|---|---|---|
| combinatorial penality | $\|Z\|_0$ | $r^*_{k,q}(Z)$ | $\mathrm{rank}(Z)$ |
| convex relaxation | $\|Z\|_1$ | **?** | $\|Z\|_*$ |

For a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, we have

| | $(1,1)$-rank | $(k,q)$-rank | $(m_1,m_2)$-rank |
|---|---|---|---|
| combinatorial penality | $\|Z\|_0$ | $r^*_{k,q}(Z)$ | $\text{rank}(Z)$ |
| convex relaxation | $\|Z\|_1$ | **?** | $\|Z\|_*$ |

- Can we define a principled relaxation of the $(k,q)$-rank?

# Atomic Norm (Chandrasekaran et al., 2012)

### Definition (Atomic norm of the set of atoms $\mathcal{A}$)

Given a set of atoms $\mathcal{A}$, the associated atomic norm is defined as

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

# Atomic Norm (Chandrasekaran et al., 2012)

### Definition (Atomic norm of the set of atoms $\mathcal{A}$)

Given a set of atoms $\mathcal{A}$, the associated atomic norm is defined as

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

NB: This is really a norm if $\mathcal{A}$ is centrally symmetric and spans $\mathbb{R}^p$

# Atomic Norm (Chandrasekaran et al., 2012)

### Definition (Atomic norm of the set of atoms $\mathcal{A}$)

Given a set of atoms $\mathcal{A}$, the associated atomic norm is defined as

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

NB: This is really a norm if $\mathcal{A}$ is centrally symmetric and spans $\mathbb{R}^p$

### Proposition (Primal and dual form of the norm)

$$\|x\|_{\mathcal{A}} \;=\; \inf\left\{\sum_{a\in\mathcal{A}} c_a \mid x = \sum_{a\in\mathcal{A}} c_a\, a, \quad c_a > 0,\, \forall a \in \mathcal{A}\right\}$$

# Atomic Norm (Chandrasekaran et al., 2012)

### Definition (Atomic norm of the set of atoms $\mathcal{A}$)

Given a set of atoms $\mathcal{A}$, the associated atomic norm is defined as

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

NB: This is really a norm if $\mathcal{A}$ is centrally symmetric and spans $\mathbb{R}^p$

### Proposition (Primal and dual form of the norm)

$$\begin{aligned} \|x\|_{\mathcal{A}} &= \inf\left\{ \sum_{a\in\mathcal{A}} c_a \;\middle|\; x = \sum_{a\in\mathcal{A}} c_a\, a, \quad c_a > 0,\, \forall a \in \mathcal{A} \right\} \\ \|x\|_{\mathcal{A}}^* &= \sup_{a\in\mathcal{A}} \langle a, x\rangle \end{aligned}$$

# Examples of atomic norms

$$\|x\|_{\mathcal{A}} = \inf \left\{ \sum_{a \in \mathcal{A}} c_a \mid x = \sum_{a \in \mathcal{A}} c_a \, a, \quad c_a > 0, \ \forall a \in \mathcal{A} \right\}$$

$$\|x\|_{\mathcal{A}}^{*} = \sup_{a \in \mathcal{A}} \langle a, x \rangle$$

# Examples of atomic norms

$$\begin{aligned}\|x\|_{\mathcal{A}} &= \inf\left\{ \sum_{a\in\mathcal{A}} c_a \;\middle|\; x = \sum_{a\in\mathcal{A}} c_a\, a, \quad c_a > 0,\ \forall a \in \mathcal{A} \right\} \\ \|x\|_{\mathcal{A}}^{*} &= \sup_{a\in\mathcal{A}} \langle a, x\rangle\end{aligned}$$

- vector $\ell_1$-norm: $x \mapsto \|x\|_1$

$$\mathcal{A} = \left\{ \pm e_k \;\middle|\; 1 \le k \le p \right\}$$

# Examples of atomic norms

$$
\begin{aligned}
\|x\|_{\mathcal{A}} &= \inf \left\{ \sum_{a\in\mathcal{A}} c_a \;\middle|\; x = \sum_{a\in\mathcal{A}} c_a\, a, \quad c_a > 0,\ \forall a \in \mathcal{A} \right\} \\
\|x\|_{\mathcal{A}}^{*} &= \sup_{a\in\mathcal{A}} \langle a, x \rangle
\end{aligned}
$$

- vector $\ell_1$-norm: $x \mapsto \|x\|_1$

$$\mathcal{A} = \left\{ \pm e_k \;\middle|\; 1 \leq k \leq p \right\}$$

- matrix trace norm: $Z \mapsto \|Z\|_*$ (sum of singular value)

$$\mathcal{A} = \left\{ ab^\top \;\middle|\; a \in \mathbb{S}^{m_1-1}, b \in \mathbb{S}^{m_2-1} \right\}$$

# Examples of atomic norms

$$\|x\|_{\mathcal{A}} = \inf \left\{ \sum_{a \in \mathcal{A}} c_a \mid x = \sum_{a \in \mathcal{A}} c_a\, a, \quad c_a > 0,\ \forall a \in \mathcal{A} \right\}$$

$$\|x\|_{\mathcal{A}}^{*} = \sup_{a \in \mathcal{A}} \langle a, x \rangle$$

- vector $\ell_1$-norm: $x \mapsto \|x\|_1$

$$\mathcal{A} = \left\{ \pm e_k \mid 1 \leq k \leq p \right\}$$

- matrix trace norm: $Z \mapsto \|Z\|_*$ (sum of singular value)

$$\mathcal{A} = \left\{ ab^\top \mid a \in \mathbb{S}^{m_1 - 1}, b \in \mathbb{S}^{m_2 - 1} \right\}$$

# A convex relaxation of the $(k, q)$-rank

With

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

# A convex relaxation of the $(k, q)$-rank

With

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

consider the set of atoms

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

# A convex relaxation of the $(k, q)$-rank

With

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

consider the set of atoms

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

The atomic norm associated with $\mathcal{A}_{k,q}$ is

$$\Omega_{k,q}(Z) = \inf \left\{ \sum_{A \in \mathcal{A}_{k,q}} c_A \ \middle|\ Z = \sum_{A \in \mathcal{A}_{k,q}} c_A\, A, \quad c_A > 0,\ \forall A \in \mathcal{A} \right\}$$

# A convex relaxation of the $(k, q)$-rank

With

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

consider the set of atoms

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

The atomic norm associated with $\mathcal{A}_{k,q}$ is

$$\Omega_{k,q}(Z) = \inf\left\{ \sum_{A \in \mathcal{A}_{k,q}} c_A \ \Big| \ Z = \sum_{A \in \mathcal{A}_{k,q}} c_A\, A, \quad c_A > 0,\ \forall A \in \mathcal{A} \right\}$$

so that

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

Call $\Omega_{k,q}$ the $(k, q)$-trace norm and solutions the $(k, q)$-sparse SVDs.

# Properties of the $(k, q)$-trace norm

## Nesting property

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

# Properties of the $(k, q)$-trace norm

## Nesting property

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

## Dual norm and reformulation

- Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm.
- Let $\mathcal{G}_{k,q} = \{(I, J) \subset [\![1, m_1]\!] \times [\![1, m_2]\!],\ |I| = k, |J| = q\}$

# Properties of the $(k, q)$-trace norm

## Nesting property

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

## Dual norm and reformulation

- Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm.
- Let $\mathcal{G}_{k,q} = \big\{(I, J) \subset [\![1, m_1]\!] \times [\![1, m_2]\!],\ |I| = k, |J| = q\big\}$

Given that $\|x\|_{\mathcal{A}}^* = \sup_{a\in\mathcal{A}} \langle a, x\rangle$, we have

$$\Omega_{k,q}^*(Z) = \max_{(I,J)\in\mathcal{G}_{k,q}} \|Z_{I,J}\|_{\mathrm{op}} \qquad\qquad \text{and}$$

# Properties of the $(k, q)$-trace norm

## Nesting property

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

## Dual norm and reformulation

- Let $\|\cdot\|_{\mathrm{op}}$ denote the operator norm.
- Let $\mathcal{G}_{k,q} = \{(I,J) \subset [\![1, m_1]\!] \times [\![1, m_2]\!],\ |I| = k, |J| = q\}$

Given that $\|x\|^*_{\mathcal{A}} = \sup_{a \in \mathcal{A}} \langle a, x \rangle$, we have

$$\Omega^*_{k,q}(Z) = \max_{(I,J)\in\mathcal{G}_{k,q}} \|Z_{I,J}\|_{\mathrm{op}} \qquad \text{and}$$

$$\Omega_{k,q}(Z) = \inf\left\{ \sum_{(I,J)\in\mathcal{G}_{k,q}} \left\|A^{(IJ)}\right\|_* \ :\ Z = \sum_{(I,J)\in\mathcal{G}_{k,q}} A^{(IJ)}\ ,\ \mathrm{supp}(A^{(IJ)}) \subset I \times J \right\}$$

# The (k,q)-CUT-norm: an $\ell_\infty$ counterpart

With the following subset of $\mathcal{A}_k^m$:

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

# The (k,q)-CUT-norm: an $\ell_\infty$ counterpart

With the following subset of $\mathcal{A}_k^m$:

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

consider the set of atoms

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

# The (k,q)-CUT-norm: an $\ell_\infty$ counterpart

With the following subset of $\mathcal{A}_k^m$:

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

consider the set of atoms

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

- Denote $\widetilde{\Omega}_{k,q}$ the "$\ell_\infty$" counterpart of the $(k,q)$-trace norm.

# The (k,q)-CUT-norm: an $\ell_\infty$ counterpart

With the following subset of $\mathcal{A}_k^m$:

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

consider the set of atoms

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

- Denote $\widetilde{\Omega}_{k,q}$ the "$\ell_\infty$" counterpart of the $(k,q)$-trace norm.
- When $k = m_1$ and $q = m_2$, $\widetilde{\Omega}_{n,n}$ is the gauge function of the CUT-polytope of a bipartite graph (Deza and Laurent, 1997).

# Vector case

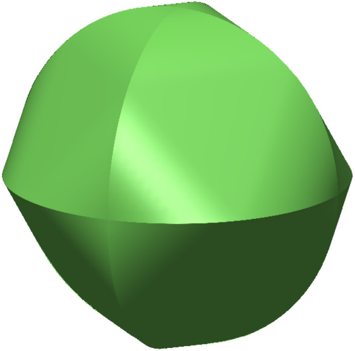
When $q = m_2 = 1$, we retrieve vector norms:

- $\Omega_{k,1} = \theta_k$ is the $k$-support norm of Argyriou et al. (2012).
- $\widetilde{\Omega}_{k,1} = \kappa_k$ with $\kappa_k$ the vector Ky Fan norm.

# Vector case

When $q = m_2 = 1$, we retrieve vector norms:

- $\Omega_{k,1} = \theta_k$ is the $k$-support norm of Argyriou et al. (2012).
- $\widetilde{\Omega}_{k,1} = \kappa_k$ with $\kappa_k$ the vector Ky Fan norm.

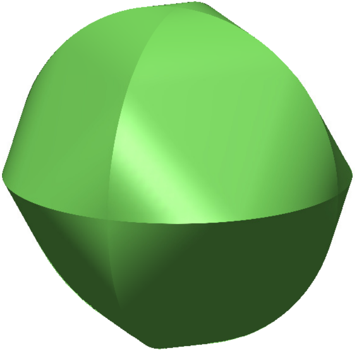


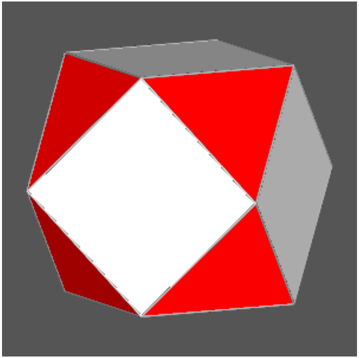
# Vector case

When $q = m_2 = 1$, we retrieve vector norms:

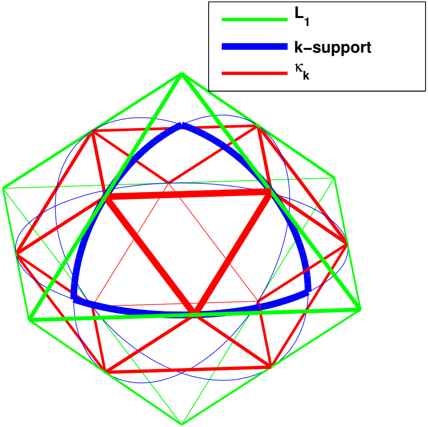
- $\Omega_{k,1} = \theta_k$ is the $k$-support norm of Argyriou et al. (2012).
- $\widetilde{\Omega}_{k,1} = \kappa_k$ with $\kappa_k$ the vector Ky Fan norm.

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





# Relation between unit balls



$\theta_k$, $\kappa_k$ and $\frac{1}{\sqrt{k}}\|\cdot\|_1$ for $k = 2$ in $\mathbb{R}^3$.

# Learning matrices with sparse factors

## Sparse bilinear regression

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

# Learning matrices with sparse factors

## Sparse bilinear regression

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

## Subspace clustering

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

# Learning matrices with sparse factors

## Sparse bilinear regression

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

## Subspace clustering

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

## Rank $r$ sparse PCA

$$\min_Z \frac{1}{2}\|\hat{\Sigma}_n - Z\|_F^2 \quad + \quad \lambda\,\Omega_{k,k}(Z) \quad \text{s.t.} \quad Z \succeq 0\,, \qquad \text{or}$$

$$\min_Z \frac{1}{2}\|\hat{\Sigma}_n - Z\|_F^2 \quad + \quad \lambda\,\Omega_{k,\succeq}(Z)\,,$$

with $\Omega_{k,\succeq}$ the atomic norm for the set $\mathcal{A}_{k,\succeq} = \{aa^\top, a \in \mathcal{A}_k\}$.

# Statistical guarantees

# Statistical dimension (Amelunxen et al., 2013)

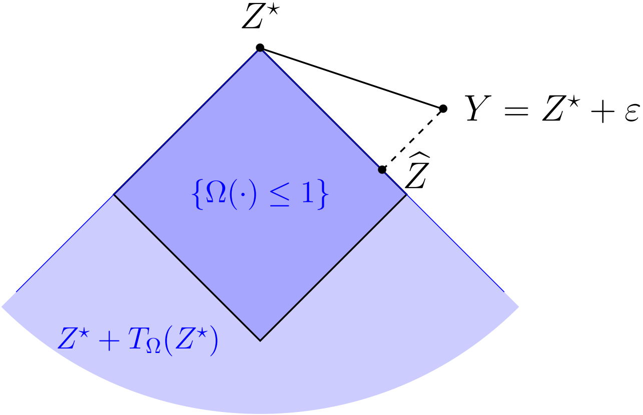

$Z^\star$
$Y = Z^\star + \varepsilon$
$\widehat{Z}$
$\{\Omega(\cdot) \leq 1\}$
$Z^\star + T_\Omega(Z^\star)$

figure inspired by Amelunxen et al. (2013)

# Statistical dimension (Amelunxen et al., 2013)

Tangent cone:

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

# Statistical dimension (Amelunxen et al., 2013)

Tangent cone:

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

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

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

where

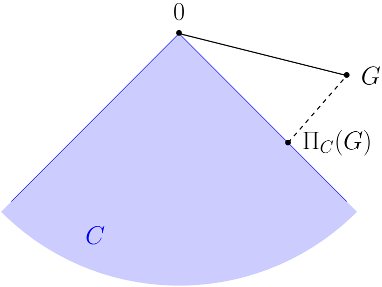
- $G$ is a matrix with i.i.d. standard normal entries
- $\Pi_{T_\Omega(Z)}(G)$ is the orthogonal projection of $G$ onto $T_\Omega(Z)$.

"The statistical dimension $\delta$ is the unique continuous, rotation-invariant localization valuation on the set of convex cones that satisfies $\delta(L) = \dim(L)$ for any subspace $L$."

# Relation between Gaussian width and statistical dimension

Gaussian width of intersection of the cone with a Euclidean ball:

$$
\begin{aligned}
w(C) &= \max_{U \in T_\Omega(Z) \cap \mathbb{S}^{d-1}} \langle U, G \rangle \\
&= \mathbb{E}\big[\, \|\Pi_C(G)\|_{\mathrm{Fro}} \,\big].
\end{aligned}
$$



Amelunxen et al. (2013) show that

$$w(C)^2 \le \mathfrak{S}(C) \le w(C)^2 + 1.$$

# General Null Space property

Consider the optimization problem

$$\min_{Z} \Omega(Z) \qquad \text{s.t.} \qquad y = \mathcal{X}(Z) \tag{1}$$

# General Null Space property

Consider the optimization problem

$$\min_{Z} \Omega(Z) \qquad \text{s.t.} \qquad y = \mathcal{X}(Z) \tag{1}$$

### Theorem (NSP)

*$Z^*$ is the unique optimal solution of (1) if and only if*

$$\mathrm{Ker}(\mathcal{X}) \cap T_{\Omega}(Z^*) = \varnothing.$$

# General Null Space property

Consider the optimization problem

$$\min_{Z} \Omega(Z) \qquad \text{s.t.} \qquad y = \mathcal{X}(Z) \tag{1}$$

### Theorem (NSP)

*$Z^*$ is the unique optimal solution of (1) if and only if*

$$\mathrm{Ker}(\mathcal{X}) \cap T_{\Omega}(Z^*) = \varnothing.$$

Note: this motivates a posteriori the construction of atomic norms.

# Nullspace property and $\mathfrak{S}$ (Chandrasekaran et al., 2012)

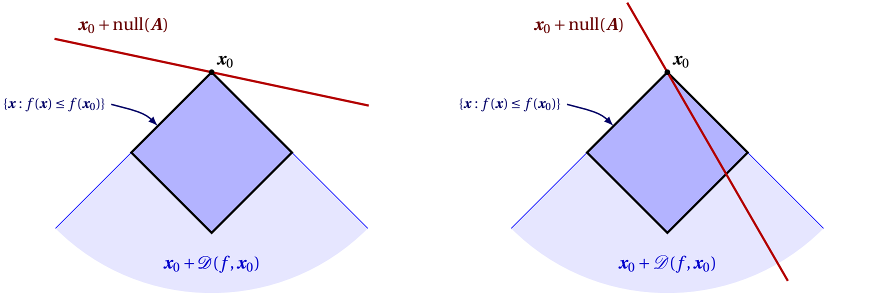

$x_0 + \mathrm{null}(A)$
$x_0$
$\{x : f(x) \leq f(x_0)\}$
$x_0 + \mathscr{D}(f, x_0)$
$x_0 + \mathrm{null}(A)$
$x_0$
$\{x : f(x) \leq f(x_0)\}$
$x_0 + \mathscr{D}(f, x_0)$

Figure from Amelunxen et al. (2013)

## Exact recovery from random measurements

With $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$ rand. lin. map from the std Gaussian ensemble

$$\widehat{Z} = \arg\min_{Z} \Omega(Z) \quad \text{s.th.} \quad \mathcal{X}(Z) = y$$

is equal to $Z^\star$ w.h.p. as soon as $\quad n \geq \mathfrak{S}(Z^\star, \Omega)$.

# Denoising with an atomic norm

If

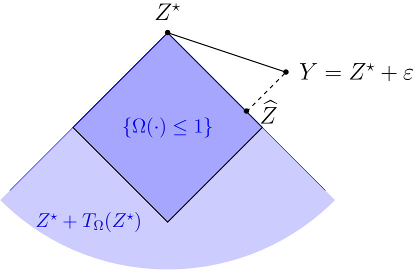
- $Y = Z^\star + \frac{\sigma}{\sqrt{n}}\epsilon$
- with $\epsilon$ standard Gaussian,

then

$$\widehat{Z} = \arg\min_{Z} \|Z - Y\|_{\mathrm{Fro}} \quad \text{s.t.} \quad \Omega(Z) \le \Omega(Z^\star)$$

satisfies

$$\boxed{\mathbb{E}\|\widehat{Z} - Z^\star\|^2 \le \frac{\sigma^2}{n}\mathfrak{S}(Z^\star, \Omega).}$$


$Z^\star$
$Y = Z^\star + \varepsilon$
$\widehat{Z}$
$\{\Omega(\cdot) \le 1\}$
$Z^\star + T_\Omega(Z^\star)$

# Statistical dimension at elements of $\widetilde{\mathcal{A}}_{k,q}$

- Consider an element $ab^\top \in \widetilde{\mathcal{A}}_{k,q}$.
- We have $\|ab^\top\|_0 = kq$.

So

| Matrix norm | $\mathfrak{S}$ |
|---|---|
| $\ell_1$ | $\Theta(kq \ \log \frac{m_1 m_2}{kq})$ |
| trace-norm | $\Theta(m_1 + m_2)$ |
| $\ell_1$ + trace-norm | **?** |
| $(k, q)$-trace | **?** |
| $(k, q)$-cut | **?** |

# Theoretical results

Proposition (UB on $\mathfrak{S}(A, \widetilde{\Omega}_{k,q})$)

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

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

# Theoretical results

## Proposition (UB on $\mathfrak{S}(A, \widetilde{\Omega}_{k,q})$)

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

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

## Proposition (UB on $\mathfrak{S}(A, \Omega_{k,q})$)

*Let $A = ab^\top \in \mathcal{A}_{k,q}$ with $I_0 = supp(a)$ and $J_0 = supp(b)$.*

$$\textit{Let} \quad \gamma(a,b) := (k \min_{i \in I_0} a_i^2) \wedge (q \min_{j \in J_0} b_j^2),$$

*we have*

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

# Summary of results for statistical dimension

| Matrix norm | $\mathfrak{S}$ | Vector norm | $\mathfrak{S}$ |
|---|---|---|---|
| $\ell_1$ | $\Theta(kq\ \log \frac{m_1 m_2}{kq})$ | $\ell_1$ | $\Theta(k \log \frac{p}{k})$ |
| trace-norm | $\Theta(m_1 + m_2)$ | $\ell_2$ | $p$ |
| $\ell_1 +$ trace-n. | $\Omega\big(kq \wedge (m_1 + m_2)\big)$[1] | elastic net | $\Theta(k \log \frac{p}{k})$ |
| $(k, q)$-trace | $\mathcal{O}((k \vee q) \log (m_1 \vee m_2))$ | $k$-support | $\Theta(k \log \frac{p}{k})$ |
| $(k, q)$-cut | $\mathcal{O}(k \log \frac{m_1}{k} + q \log \frac{m_2}{q})$ | $\kappa_k$ | $\Theta(k \log \frac{p}{k})$ |
| "cut-norm" | $\mathcal{O}(m_1 + m_2)$ | $\ell_\infty$ | $p$ |

[1] Proof relying on a result of Oymak et al. (2012)

# Algorithm

# Working set algorithm

Given a working set $\mathcal{S}$ of blocks $(I, J)$, solve the restricted problem

$$\min_{Z,\,(A^{(IJ)})_{(I,J)\in\mathcal{S}}} \quad \mathcal{L}(Z) + \lambda \sum_{(I,J)\in\mathcal{S}} \left\|A^{(IJ)}\right\|_*$$

$$Z = \sum_{(I,J)\in\mathcal{S}} A^{(IJ)}\ ,\ \operatorname{supp}(A^{(IJ)}) \subset I \times J.$$

## Proposition

*The global problem is solved by a solution $Z_{\mathcal{S}}$ of the restricted problem if and only if*

$$\forall (I, J) \in \mathcal{G}_{k,q}, \quad \left\|\left[\nabla\mathcal{L}(Z_{\mathcal{S}})\right]_{I,J}\right\|_{\mathrm{op}} \leq \lambda. \qquad (\star)$$

# Working set algorithm

## Active set algorithm

Iterate:

1. Solve the restricted problem
2. Look for $(I, J)$ that violates $(\star)$
   - If none exists, terminate the algorithm !
   - Else add the found $(I, J)$ to $\mathcal{S}$

**Problem**: step 2 require to solve a rank-1 SPCA problem → NP-hard

**Idea:** Leverage the work on algorithms that attempt to solve rank-1 SPCA like

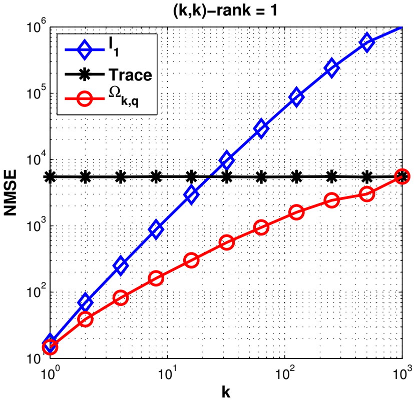
- convex relaxations,
- truncated power iteration method

to heuristically find blocks potentially violating the constraint.

# Experiments

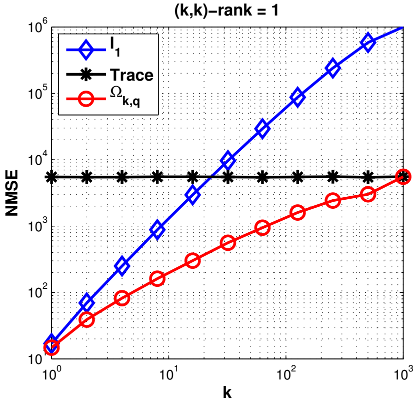
# Denoising results

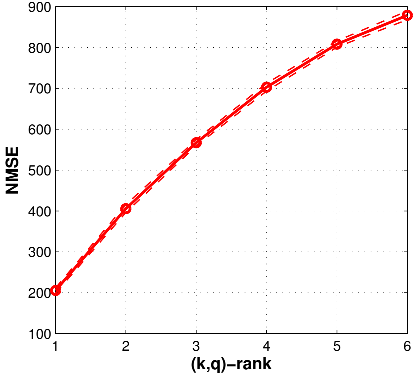
- $Z \in \mathbb{R}^{1000\times 1000}$ with $Z = \sum_{i=1}^{r} a_i b_i^\top + \sigma G$ and $a_i b_i^\top \in \mathcal{A}_{k,q}$
- $k = q$
- $\sigma^2$ small $\Rightarrow$ MSE $\propto \mathfrak{S}(ab^\top, \Omega_{k,q})\, \sigma^2$



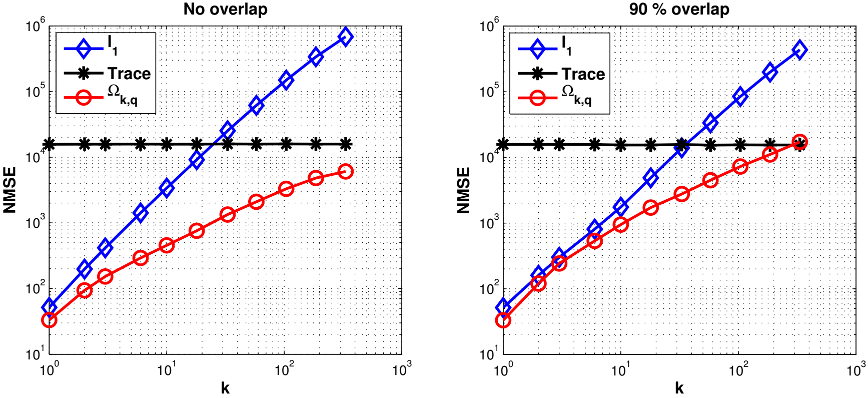
# Denoising results

- $Z \in \mathbb{R}^{1000\times 1000}$ with $Z = \sum_{i=1}^{r} a_i b_i^\top + \sigma G$ and $a_i b_i^\top \in \mathcal{A}_{k,q}$
- $k = q$
- $\sigma^2$ small $\Rightarrow$ MSE $\propto \mathfrak{S}(ab^\top, \Omega_{k,q})\, \sigma^2$

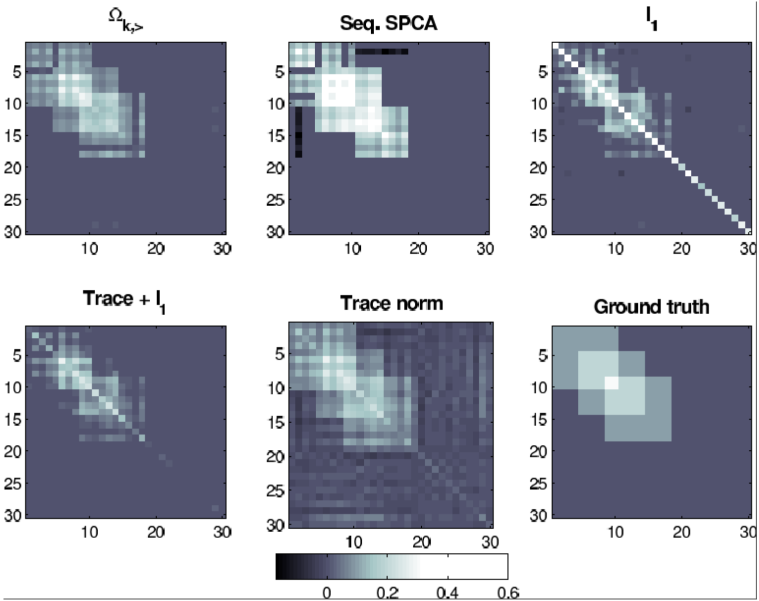




# Denoising results $[Z \in \mathbb{R}^{300\times 300}$ and $\sigma^2$ small $\Rightarrow$ MSE $\propto \mathfrak{S}(ab^\top, \Omega_{k,q})\,\sigma^2]$



# Empirical results for sparse PCA



# Conclusions

## Summary

- Gain in statistical performance at the expense of theoretical tractability.
- Even though the problem is NP-hard the structure of the convex problem can be exploited to devise efficient heuristics.

## Not discussed

- slow rate analysis
- purely geometric results

## Open questions and future work

- Generalization to the case where $(k, q)$ can be different for each pair of factors and not known a priori.

# References I

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2013). Living on the edge: Phase transitions in convex programs with random data. Technical Report 1303.6672, arXiv.

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Stat.*, 37(5B):2877–2921.

Argyriou, A., Foygel, R., and Srebro, N. (2012). Sparse prediction with the $k$-support norm. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Adv. Neural. Inform. Process Syst.*, volume 25, pages 1457–1465. Curran Associates, Inc.

Bach, F. (2013). Convex relaxations of structured matrix factorizations. Technical Report 1309.3117, arXiv.

Bach, F., Mairal, J., and Ponce, J. (2008). Convex sparse matrix factorizations. Technical Report 0812.1869, arXiv.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849.

d'Aspremont, A., Bach, F., and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9:1269–1294.

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

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

# References II

Doan, X. V. and Vavasis, S. A. (2013). Finding approximately rank-one submatrices with the nuclear norm and $\ell_1$ norms. *SIAM J. Optimiz.*, 23(4):2502–2540.

Elad, M. and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745.

Jameson, G. J. O. (1987). *Summing and Nuclear Norms in Banach Space Theory*. Number 8 in London Mathematical Society Student Texts. Cambridge University Press.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, 11:517–553.

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

Oymak, S., Jalali, A., Fazel, M., Eldar, Y. C., and Hassibi, B. (2012). Simultaneously structured models with application to sparse and low-rank matrices. Technical Report 1212.3753, arXiv.

Richard, E., Savalle, P.-A., and Vayatis, N. (2012). 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.

Wang, Y.-X., Xu, H., and Leng, C. (2013). Provable subspace clustering: When LRR meets SSC. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Adv. Neural. Inform. Process Syst.*, volume 26, pages 64–72. Curran Associates, Inc.

# References III

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Yuan, X.-T. and Zhang, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, 14:889–925.