New matrix norms for structured matrix estimation

Jean-Philippe Vert

Sparse Representations, Numerical Linear Algebra, and Optimization workshop
BIRS, Banff, Oct 5-10, 2014
Outline

1 Atomic norms
2 Sparse matrices with disjoint column supports
3 Low-rank matrices with sparse factors
Outline

1. Atomic norms

2. Sparse matrices with disjoint column supports

3. Low-rank matrices with sparse factors
Atomic Norm  (Chandrasekaran et al., 2012)

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

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

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

Primal and dual form of the norm

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

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

- **Vector $\ell_1$-norm:** $x \in \mathbb{R}^p \mapsto \|x\|_1$

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

- **Matrix trace norm:** $Z \in \mathbb{R}^{m_1 \times m_2} \mapsto \|Z\|_*$ (sum of singular value)

  $\mathcal{A} = \{ ab^\top : a \in \mathbb{R}^{m_1}, b \in \mathbb{R}^{m_2}, \|a\|_2 = \|b\|_2 = 1 \}$
Group lasso (Yuan and Lin, 2006)

For $x \in \mathbb{R}^p$ and $\mathcal{G} = \{g_1, \ldots, g_G\}$ a partition of $[1, p]$:

$$\|x\|_{1,2} = \sum_{g \in \mathcal{G}} \|x_g\|_2$$

is the atomic norm associated to the set of atoms

$$\mathcal{A}_\mathcal{G} = \bigcup_{g \in \mathcal{G}} \{u \in \mathbb{R}^p : \text{supp}(u) = g, \|u\|_2 = 1\}$$

$$\mathcal{G} = \{\{1, 2\}, \{3\}\}$$

$$\|x\|_{1,2} = \|(x_1, x_2)^\top\|_2 + \|x_3\|_2$$

$$= \sqrt{x_1^2 + x_2^2} + \sqrt{x_3^2}$$
Group lasso with overlaps

How to generalize the group lasso when the groups overlap?

- Set features to zero by groups (Jenatton et al., 2011)
  \[ \| x \|_{1,2} = \sum_{g \in G} \| x_g \|_2 \]

- Select support as a union of groups (Jacob et al., 2009)
  \[ \| x \|_{A_G} , \]
  see also MKL (Bach et al., 2004)

\[ G = \{ \{1, 2\}, \{2, 3\} \} \]
Outline

1. Atomic norms
2. Sparse matrices with disjoint column supports
3. Low-rank matrices with sparse factors
Joint work with...

Kevin Vervier, Pierre Mahé, Jean-Baptiste Veyrieras (Biomerieux)

Alexandre d’Aspremont (CNRS/ENS)
Motivation: multiclass or multitask classification problems where we want to select features specific to each class or task.

Example: recognize identify and emotion of a person from an image (Romera-Paredes et al., 2012), or hierarchical coarse-to-fine classifier (Xiao et al., 2011; Hwang et al., 2011).
From disjoint supports to orthogonal columns

Two vectors \( v_1 \) and \( v_2 \) have disjoint support iff \( |v_1| \) and \( |v_2| \) are orthogonal.

If \( \Omega_{\text{ortho}}(X) \) is a norm to estimate matrices with orthogonal columns, then

\[
\Omega_{\text{disjoint}}(X) = \Omega_{\text{ortho}}(|X|) = \min_{-W \leq X \leq W} \Omega_{\text{ortho}}(W)
\]

is a norm to estimate matrices with disjoint column supports.

How to estimate matrices with orthogonal columns?

*NOTE: more general than orthogonal matrices*
Penalty for orthogonal columns

- For $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ we want

$$x_i^\top x_j = 0 \quad \text{for} \quad i \neq j$$

- A natural "relaxation":

$$\Omega(X) = \sum_{i \neq j} \left| x_i^\top x_j \right|$$

- But not convex
Convex penalty for orthogonal columns

\[ \Omega_K(X) = \sum_{i=1}^{p} K_{ii} \| x_i \|^2 + \sum_{i \neq j} K_{ij} \left| x_i^\top x_j \right| \]

Theorem (Xiao et al., 2011)
If \( \bar{K} \) is positive semidefinite, then \( \Omega_K \) is convex, where

\[ \bar{K}_{ij} = \begin{cases} 
|K_{ii}| & \text{if } i = j, \\
-|K_{ij}| & \text{otherwise.}
\end{cases} \]
Can we be tighter?

\[ \Omega_K(X) = \sum_{i=1}^{p} \| x_i \|^2 + \sum_{i \neq j} K_{ij} \left| x_i^\top x_j \right| \]

- Let \( \mathcal{O} \) be the set of matrices of unit Frobenius norm, with orthogonal columns

\[ \mathcal{O} = \left\{ X \in \mathbb{R}^{n \times p} : X^\top X \text{ is diagonal and } \text{Trace}(X^\top X) = 1 \right\} \]

- Note that \( \forall X \in \mathcal{O}, \quad \Omega_K(X) = 1 \)

- The atomic norm \( \| X \|_\mathcal{O} \) associated to \( \mathcal{O} \) is the tightest convex penalty to recover the atoms in \( \mathcal{O} \)!
Optimality of $\Omega_K$ for $p = 2$

Theorem (Vervier, Mahé, d’Aspremont, Veyrieras and V., 2014)

For any $X \in \mathbb{R}^{n \times 2}$,

$$\Omega_K(X) = \| X \|_O^2$$

with

$$K = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$
Case $p > 2$

- $\Omega_K(X) \neq \| X \|_O^2$
- But sparse combinations of matrices in $\mathcal{O}$ may not be interesting anyway...

**Theorem (Vervier et al., 2014)**

For any $p \geq 2$, let $K$ be a symmetric $p$-by-$p$ matrix with non-negative entries and such that,

$$\forall i = 1, \ldots, p \quad K_{ii} = \sum_{j \neq i} K_{ij}.$$  

Then

$$\Omega_K(X) = \sum_{i<j} K_{ij} \| (x_i, x_j) \|_O^2.$$
Simulations

Regression $Y = XW + \epsilon$, $W$ has disjoint column support, $n = p = 10$
Example: multiclass classification of MS spectra

Features

Spectra

0
BAC
LIS
CLO
STR
CIT
ENT
ESH−SHG
YER
HAE
multi

Features
Outline

1. Atomic norms
2. Sparse matrices with disjoint column supports
3. Low-rank matrices with sparse factors
Joint work with...

Emile Richard (Stanford)

Guillaume Obozinski (Ecole des Ponts - ParisTech)
Low-rank matrices with sparse factors

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

Factors not orthogonal a priori

≠ from assuming the SVD of \( X \) is sparse
Dictionary Learning

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

- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)
Dictionary Learning / Sparse PCA

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

Dictionary Learning

\[X^T = DA\]

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

Sparse PCA

\[X^T = DA\]

- 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, 
  \[ X^\top = \begin{bmatrix} X_1^\top & \ldots & X_K^\top \end{bmatrix} \]
  with \( X_k \) low rank, so that there exists a low rank matrix \( Z_k \) such that \( X_k = Z_k X_k \). Finally,
  \[ X = ZX \quad \text{with} \quad Z = BkDiag(Z_1, \ldots, Z_K). \]

- **Sparse PCA from \( \hat{\Sigma}_n \)**

- **Sparse bilinear regression**

  \[ y = x^\top Mx' + \varepsilon \]
Existing approaches

- Bi-convex formulations

\[
\min_{U,V} \mathcal{L}(UV^T) + \lambda (\|U\|_1 + \|V\|_1),
\]

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.
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.
The \((k, q)\)-rank of a matrix

- **Sparse unit vectors:**
  \[
  \mathcal{A}_j^n = \{ a \in \mathbb{R}^n : \|a\|_0 \leq j, \|a\|_2 = 1 \}
  \]

- **\((k, q)\)-rank of a \(m_1 \times m_2\) matrix \(Z\):**
  \[
  r_{k, q}(Z) = \min \left\{ r : Z = \sum_{i=1}^{r} c_i a_i b_i^\top, (a_i, b_i, c_i) \in \mathcal{A}_{k}^{m_1} \times \mathcal{A}_{q}^{m_2} \times \mathbb{R}_+ \right\}
  \]
  \[
  = \min \left\{ \|c\|_0 : Z = \sum_{i=1}^{\infty} c_i a_i b_i^\top, (a_i, b_i, c_i) \in \mathcal{A}_{k}^{m_1} \times \mathcal{A}_{q}^{m_2} \times \mathbb{R}_+ \right\}
  \]

\[
Z = \begin{pmatrix}
\text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} \\
\text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} \\
\text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} & \text{\textbullet} \\
\end{pmatrix}
\]

\[
r_{k, q}(Z) = 3
\]
For a matrix $Z \in \mathbb{R}^{m_1 \times m_2}$, we have

<table>
<thead>
<tr>
<th>Penality Type</th>
<th>$|Z|_0$</th>
<th>rank($Z$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>combinatorial penalty</td>
<td>$|Z|_0$</td>
<td>rank($Z$)</td>
</tr>
<tr>
<td>convex relaxation</td>
<td>$|Z|_1$</td>
<td>$|Z|_\ast$</td>
</tr>
</tbody>
</table>
The \((k, q)\) trace norm (Richard et al., 2014)

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

<table>
<thead>
<tr>
<th></th>
<th>(1, 1)-rank</th>
<th>((k, q))-rank</th>
<th>((m_1, m_2))-rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>combinatorial penalty</td>
<td>(|Z|_0)</td>
<td>(r_{k,q}(Z))</td>
<td>(\text{rank}(Z))</td>
</tr>
<tr>
<td>convex relaxation</td>
<td>(|Z|_1)</td>
<td></td>
<td>(|Z|_*)</td>
</tr>
</tbody>
</table>
The \((k, q)\) trace norm (Richard et al., 2014)

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

<table>
<thead>
<tr>
<th></th>
<th>((1, 1))-rank</th>
<th>((k, q))-rank</th>
<th>((m_1, m_2))-rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>combinatorial penalty</td>
<td>(|Z|_0)</td>
<td>(r_{k,q}(Z))</td>
<td>(\text{rank}(Z))</td>
</tr>
<tr>
<td>convex relaxation</td>
<td>(|Z|_1)</td>
<td>(\Omega_{k,q}(Z))</td>
<td>(|Z|_*)</td>
</tr>
</tbody>
</table>

The \((k, q)\) trace norm \(\Omega_{k,q}(Z)\) is the atomic norm associated with

\[
\mathcal{A}_{k,q} := \left\{ ab^\top \mid a \in \mathcal{A}_{k}^{m_1}, \ b \in \mathcal{A}_{q}^{m_2} \right\},
\]

namely:

\[
\Omega_{k,q}(Z) = \inf \left\{ \|c\|_1 : Z = \sum_{i=1}^{\infty} c_i a_i b_i^\top, \ (a_i, b_i, c_i) \in \mathcal{A}_{k}^{m_1} \times \mathcal{A}_{q}^{m_2} \times \mathbb{R}_+ \right\}
\]
Some 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 \(\| \cdot \|_{op}\) 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

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

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

When $q = m_2 = 1$, $\Omega_{k,1}(x)$ is the $k$-support norm of Argyriou et al. (2012), i.e., the overlapping group lasso with all groups of size $k$. 
Statistical dimension (Amelunxen et al., 2013)

\[ Y = Z^* + \varepsilon \]

\[ \hat{Z} \]

\[ \{ \Omega(\cdot) \leq 1 \} \]

\[ Z^* \]

\[ Z^* + T_{\Omega}(Z^*) \]

\[ \mathcal{G}(Z, \Omega) := \mathbb{E} \left[ \left\| \Pi_{T_{\Omega}(Z)}(G) \right\|_{\text{Fro}}^2 \right] , \]
Exact recovery from random measurements
With $\chi : \mathbb{R}^p \rightarrow \mathbb{R}^n$ rand. lin. map from the std Gaussian ensemble

$$\hat{Z} = \arg \min_Z \Omega(Z) \quad \text{s.th.} \quad \chi(Z) = y$$

is equal to $Z^*$ w.h.p. as soon as $n \geq \mathcal{G}(Z^*, \Omega)$.
Statistical dimension of the \((k, q)\)-trace norm

**Theorem (Richard et al., 2014)**

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

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

we have

\[
\mathcal{G}(A, \Omega_{k,q}) \leq \frac{322}{\gamma^2}(k + q + 1) + \frac{160}{\gamma}(k \lor q) \log (m_1 \lor m_2) .
\]
Summary of results for statistical dimension

<table>
<thead>
<tr>
<th>Matrix norm</th>
<th>$\mathcal{S}$</th>
<th>Vector norm</th>
<th>$\mathcal{S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_1$</td>
<td>$\Theta(kq \log \frac{m_1 m_2}{kq})$</td>
<td>$\ell_1$</td>
<td>$\Theta(k \log \frac{p}{k})$</td>
</tr>
<tr>
<td>trace-norm</td>
<td>$\Theta(m_1 + m_2)$</td>
<td>$\ell_2$</td>
<td>$p$</td>
</tr>
<tr>
<td>$\ell_1 + \text{trace}$</td>
<td>$\Omega(kq \land (m_1 + m_2))$</td>
<td>elastic net</td>
<td>$\Theta(k \log \frac{p}{k})$</td>
</tr>
<tr>
<td>$(k, q)$-trace</td>
<td>$O((k \lor q) \log (m_1 \lor m_2))$</td>
<td>$k$-support</td>
<td>$\Theta(k \log \frac{p}{k})$</td>
</tr>
</tbody>
</table>

Lower bound for $\ell_1 + \text{trace}$ norm based on a result of Oymak et al. (2012)

$f = \Theta(g)$ means $(f = O(g) \& g = O(f))$

$f = \Omega(g)$ means $g = O(f)$
Working set algorithm

\[
\min_Z \mathcal{L}(Z) + \lambda \Omega_k,q(Z)
\]

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

\[
\min_{Z, (A^{(I,J)})_{(I,J)\in S}} \mathcal{L}(Z) + \lambda \sum_{(I,J)\in S} \|A^{(I,J)}\|_*
\]

\[
Z = \sum_{(I,J)\in S} A^{(I,J)} , \quad \text{supp}(A^{(I,J)}) \subset I \times J.
\]

**Proposition**

The global problem is solved by a solution \(Z_S\) of the restricted problem if and only if

\[
\forall (I, J) \in \mathcal{G}_{k,q}, \quad \left\| \left[ \nabla \mathcal{L}(Z_S) \right]_{I,J} \right\|_{\text{op}} \leq \lambda.
\]  \((\star)\)
Working set algorithm

Active set algorithm

<table>
<thead>
<tr>
<th>Iterate:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Solve the restricted problem by block coordinate descent (Tseng and Yun, 2009)</td>
</tr>
<tr>
<td>2. Look for ((I, J)) that violates ((\star))</td>
</tr>
<tr>
<td>- If none exists, terminate the algorithm !</td>
</tr>
<tr>
<td>- Else add the found ((I, J)) to (S)</td>
</tr>
</tbody>
</table>

**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.
Denoising results

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

\begin{align*}
(k,k)-\text{rank} = 1
\end{align*}

\begin{align*}
\text{NMSE} & \quad k\\
10^0 & \quad 10^1 & \quad 10^2 & \quad 10^3
\end{align*}
Denoising results

- \( Z \in \mathbb{R}^{300 \times 300} \) and \( \sigma^2 \) small \( \Rightarrow \) \( \text{MSE} \propto \mathcal{G}(ab^\top, \Omega_{k,q}) \sigma^2 \)
- \( r = 3 \) atoms, with or without overlap
Empirical results for sparse PCA

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample covariance</td>
<td>4.20 ± 0.02</td>
</tr>
<tr>
<td>Trace</td>
<td>0.98 ± 0.01</td>
</tr>
<tr>
<td>$\ell_1$</td>
<td>2.07 ± 0.01</td>
</tr>
<tr>
<td>Trace + $\ell_1$</td>
<td>0.96 ± 0.01</td>
</tr>
<tr>
<td>Sequential</td>
<td>0.93 ± 0.08</td>
</tr>
<tr>
<td>$\Omega_{k,\triangledown}$</td>
<td>0.59 ± 0.03</td>
</tr>
</tbody>
</table>
Conclusion

- Atomic norms for structured sparsity
- Gain in statistical performance at the expense of algorithmic complexity (convex but NP-hard)
- The structure of the convex problem may be exploited to devise new efficient heuristics or relaxations


