New matrix norms for structured matrix estimation

Jean-Philippe Vert

MINES ParisTech

institut Curie
Together, let's beat cancer.

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

http://www.homemade-gifts-made-easy.com/make-paper-lanterns.html

Outline

Atomic Norm (Chandrasekaran et al., 2012)

Definition

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

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

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

Primal and dual form of the norm

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

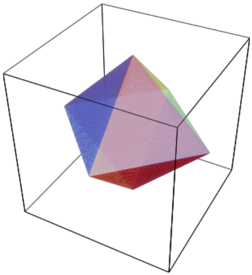
Examples

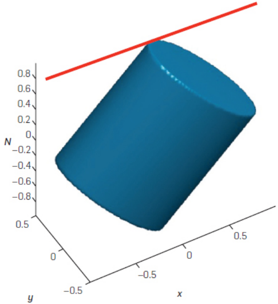
- Vector ℓ_1-norm: $x \in \mathbb{R}^p \mapsto \|x\|_1$

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

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

$$\mathcal{A} = \{ ab^\top \ : \ a \in \mathbb{R}^{m_1}, b \in \mathbb{R}^{m_2}, \| a \|_2 = \| b \|_2 = 1 \}$$





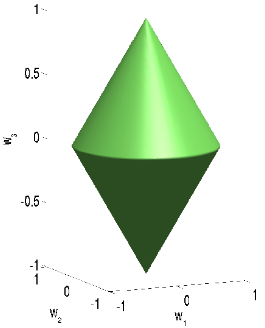
Group lasso (Yuan and Lin, 2006)

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

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



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

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

$$= \sqrt{x_1^2 + x_2^2} + \sqrt{x_3^2}$$

Group lasso with overlaps

How to generalize the group lasso when the groups overlap?

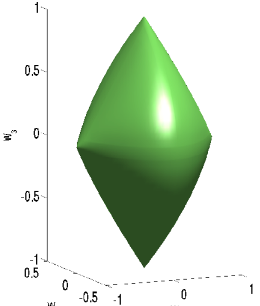
- Set features to zero by groups (Jenatton et al., 2011)

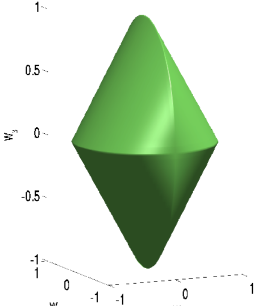
$$\| x \|_{1,2} = \sum_{g \in \mathcal{G}} \| x_g \|_2$$

- Select support as a union of groups (Jacob et al., 2009)

$$\| x \|_{\mathcal{A}_{\mathcal{G}}},$$

  see also MKL (Bach et al., 2004)





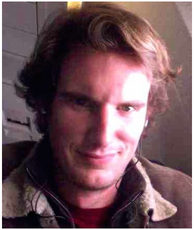
$$\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$$

Outline

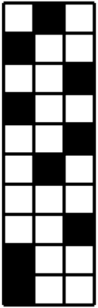
Joint work with...

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

Alexandre d'Aspremont (CNRS/ENS)

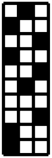


Columns with disjoint supports

$X =$ 

- 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

$X =$ 

- Two vectors v_1 and v_2 have disjoint support iff $|v_1|$ and $|v_2|$ are orthogonal
- If $\Omega_{ortho}(X)$ is a norm to estimate matrices with orthogonal columns, then

$$\Omega_{disjoint}(X) = \Omega_{ortho}(|X|) = \min_{-W \leq X \leq W} \Omega_{ortho}(W)$$

  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 Ω_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

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

- 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 Ω_K for $p = 2$

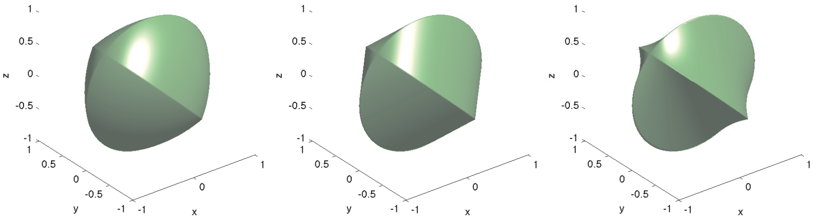
Theorem (Vervier, Mahé, d'Aspremont, Veyrieras and V., 2014)

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

$$\Omega_K(X) = \|\, X \,\|_{\mathcal{O}}^2$$

with

$$K = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} .$$



Case $p > 2$

- $\Omega_K(X) \neq \| X \|_{\mathcal{O}}^2$
- 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,

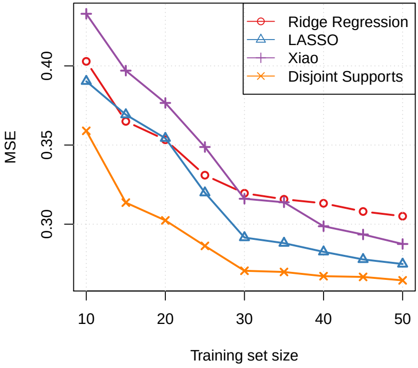
$$\forall i = 1, \ldots, p \quad K_{ii} = \sum_{j \neq i} K_{ij} \,.$$

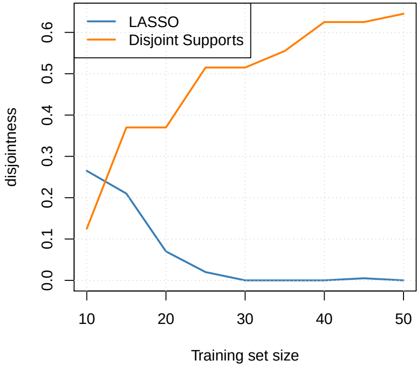
Then

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

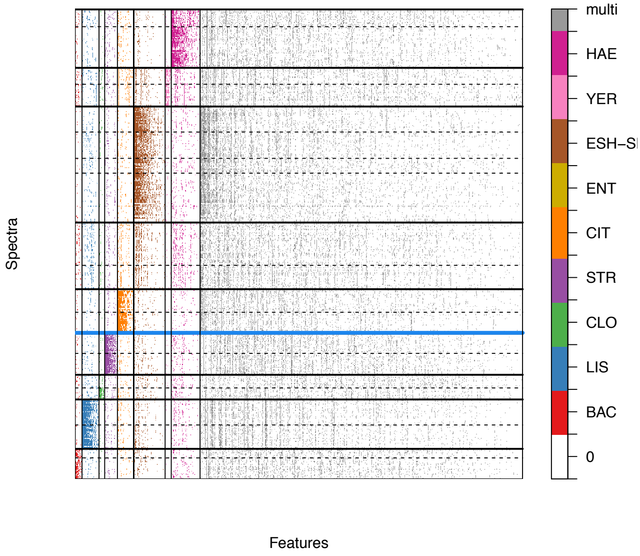
Simulations

Regression $Y = XW + \epsilon$, W has disjoint column support, $n = p = 10$





Example: multiclass classification of MS spectra



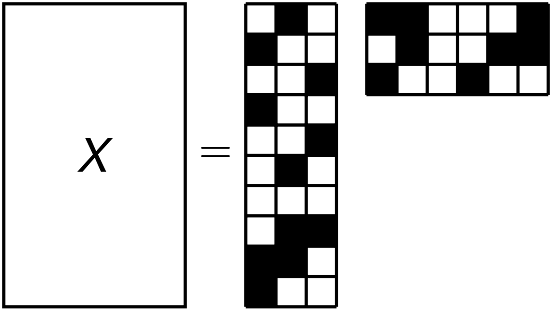
Outline

Joint work with...

Emile Richard (Stanford)

Guillaume Obozinski (Ecole des Ponts - ParisTech)

Low-rank matrices with sparse factors



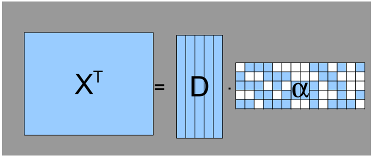
$$X = \sum_{i=1}^{r} u_i v_i^\top$$

- factors not orthogonal a priori
- $\neq$ 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

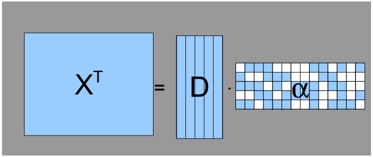


- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)

Dictionary Learning /Sparse PCA

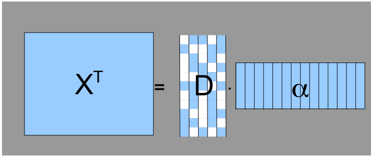
$$\min_{\substack{A \in \mathbb{R}^{k \times n} \\ D \in \mathbb{R}^{p \times k}}} \sum_{i=1}^{n} \|x_i - D\alpha_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\alpha_i\|_1 \quad \text{s.t.} \quad \forall j, \ \|d_j\|_2 \leq 1.$$

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,

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

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

- 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 ℓ_1/trace norm.

Propose algorithms based on this formulation.

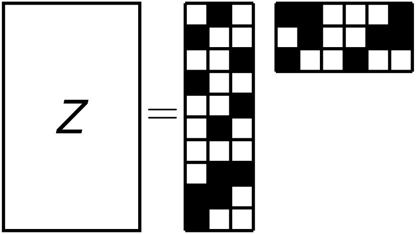
The (k, q)-rank of a matrix

- 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:

$$\begin{aligned} r_{k,q}(Z) &= \min \left\{ r \ : \ Z = \sum_{i=1}^{r} c_i a_i b_i^\top, (a_i, b_i, c_i) \in \mathcal{A}_k^{m_1} \times \mathcal{A}_q^{m_2} \times \mathbb{R}_+ \right\} \\ &= \min \left\{ \| c \|_0 \ : \ Z = \sum_{i=1}^{\infty} c_i a_i b_i^\top, (a_i, b_i, c_i) \in \mathcal{A}_k^{m_1} \times \mathcal{A}_q^{m_2} \times \mathbb{R}_+ \right\} \end{aligned}$$



$$r_{k,q}(Z) = 3$$

The (k, q) trace norm (Richard et al., 2014)

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

combinatorial penality	$\|Z\|_0$		$\text{rank}(Z)$
convex relaxation	$\|Z\|_1$		$\|Z\|_*$

The (k, q) trace norm (Richard et al., 2014)

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	$\|Z\|_1$		$\|Z\|_*$

The (k, q) trace norm (Richard et al., 2014)

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	$\|Z\|_1$	$\Omega_{k,q}(Z)$	$\|Z\|_*$

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:

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

Dual norm and reformulation

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

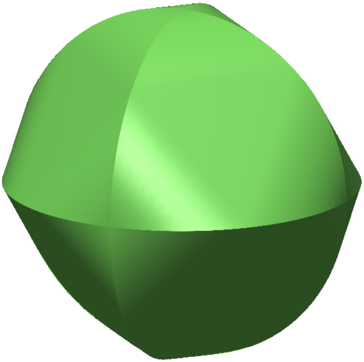
Given that $\|x\|_{\mathcal{A}}^* = \sup_{a\in\mathcal{A}} \langle a, x\rangle$, we have

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

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

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)

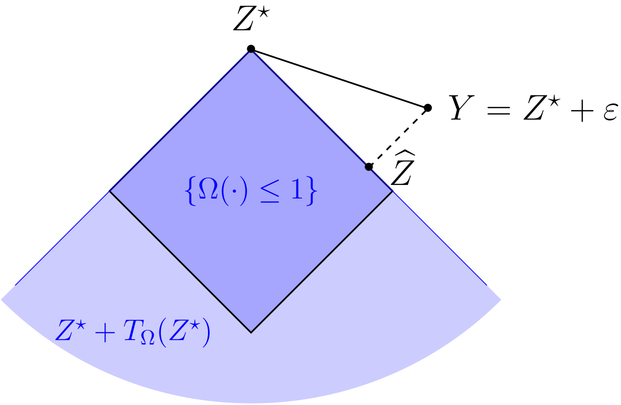


figure inspired by Amelunxen et al. (2013)

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

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

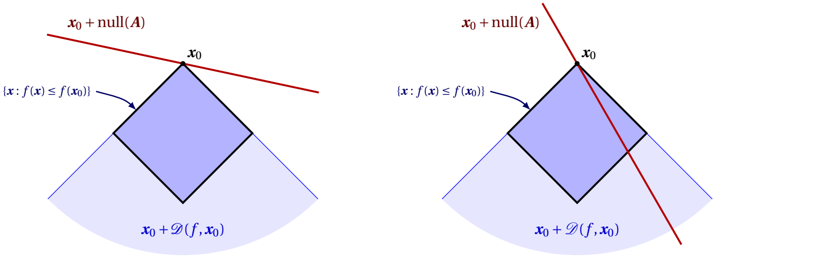


Figure from Amelunxen et al. (2013)

Exact recovery from random measurements

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

$$\widehat{Z} = \operatorname*{argmin}_Z \Omega(Z) \quad \text{s.th.} \quad \mathcal{X}(Z) = y$$

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

Statistical dimension of the (k, q)-trace norm

Theorem (Richard et al., 2014)

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

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

we have

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

Summary of results for statistical dimension

Matrix norm	$\mathfrak{S}$	Vector norm	$\mathfrak{S}$
ℓ_1	$\Theta(kq \ \log \frac{m_1 m_2}{kq})$	ℓ_1	$\Theta(k \log \frac{p}{k})$
trace-norm	$\Theta(m_1 + m_2)$	ℓ_2	p
ℓ_1 + trace	$\Omega\big(kq \wedge (m_1 + m_2)\big)$	elastic net	$\Theta(k \log \frac{p}{k})$
(k, q)-trace	$\mathcal{O}((k \vee q) \log (m_1 \vee m_2))$	k-support	$\Theta(k \log \frac{p}{k})$

Lower bound for ℓ_1+ trace norm based on a result of Oymak et al. (2012)
$f = \Theta(g)$ means $(f = \mathcal{O}(g) \& g = \mathcal{O}(f))$

$f = \Omega(g)$ means $g = \mathcal{O}(f)$

Working set algorithm

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

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 by block coordinate descent (Tseng and Yun, 2009)
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 $\rightarrow$ NP-hard

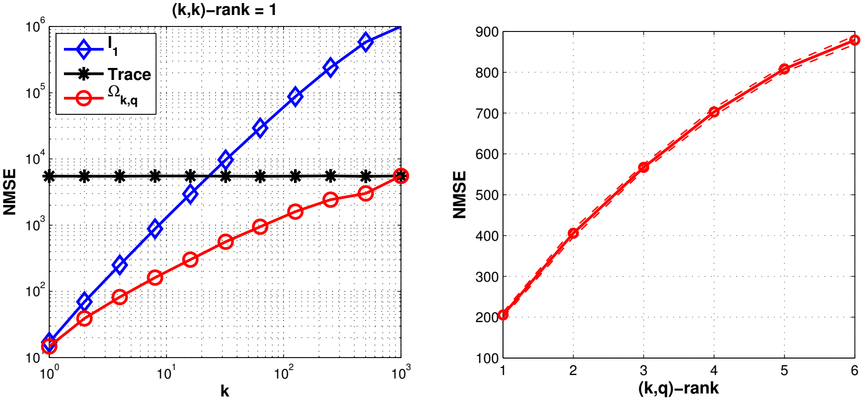
Idea: Leverage the work on algorithms that attempt to solve rank-1 SPCA like

- convex relaxations,
- truncated power iteration method

to heuristically find blocks potentially violating the constraint.

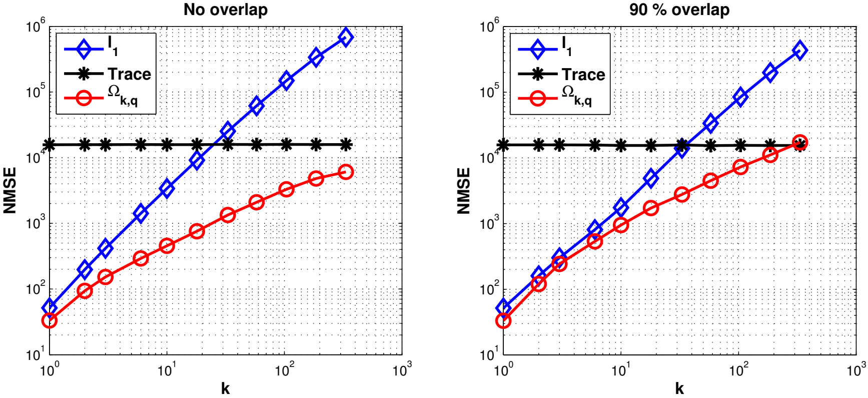
Denoising results

- $Z \in \mathbb{R}^{1000\times1000}$ with $Z = \sum_{i=1}^{r} a_i b_i^\top + \sigma G$ and $a_i b_i^\top \in \mathcal{A}_{k,q}$
- $k = q$
- σ^2 small $\Rightarrow$ MSE $\propto \mathfrak{S}(ab^\top, \Omega_{k,q})\,\sigma^2$

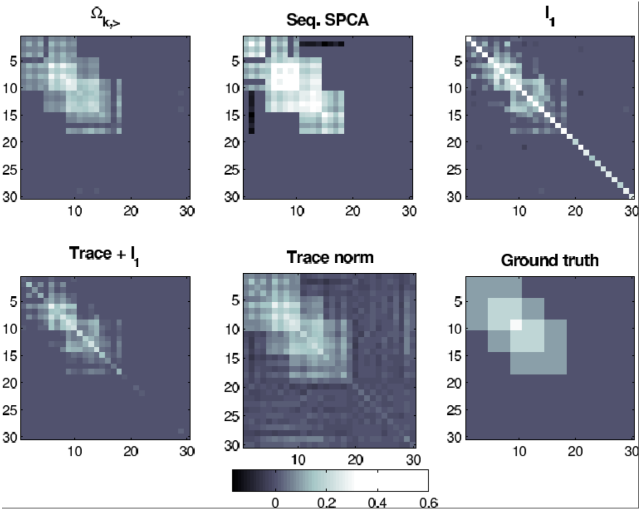


Denoising results

- $[Z \in \mathbb{R}^{300\times 300}$ and σ^2 small $\Rightarrow$ MSE $\propto \mathfrak{S}(ab^\top, \Omega_{k,q})\, \sigma^2]$
- $r = 3$ atoms, with or without overlap



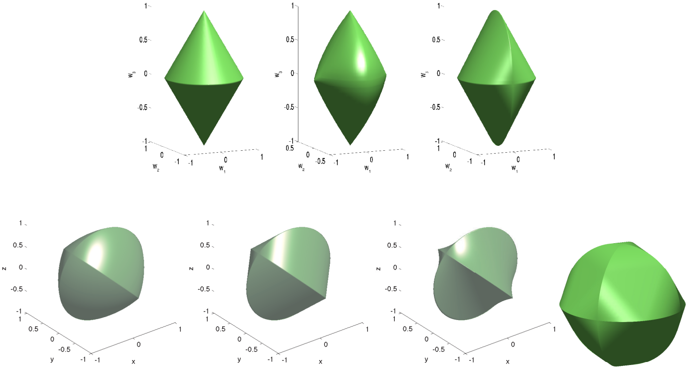
Empirical results for sparse PCA



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	$\mathbf{0.59 \pm 0.03}$

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

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

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.

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

Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 6, New York, NY, USA. ACM.

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.

Doan, X. V. and Vavasis, S. A. (2013). Finding approximately rank-one submatrices with the nuclear norm and ℓ_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 Trans. Image Process.*, 15(12):3736–3745.

Hwang, S. J. J., Grauman, K., and Sha, F. (2011). Learning a tree of metrics with disjoint visual features. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 621–629.

References II

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). 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. ACM.

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.*, 12:2777–2824.

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., Obozinski, G., and Vert, J.-P. (2014). Tight convex relaxations for sparse matrix factorization. In *Adv. Neural. Inform. Process Syst.*

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.

Romera-Paredes, B., Argyriou, A., Berthouze, N., and Pontil, M. (2012). Exploiting unrelated tasks in multi-task learning. *J. Mach. Learn. Res. - Proceedings Track,*, 22:951–959.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, 117(1-2):387–423.

Vervier, K., Mahé, P., D'Aspremont, A., Veyrieras, J.-B., and Vert, J.-P. (2014). On learning matrices with orthogonal columns or disjoint supports. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8726 of *Lecture Notes in Computer Science*, pages 274–289. Springer Berlin Heidelberg.

References III

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.

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.

Xiao, L., Zhou, D., and Wu, M. (2011). Hierarchical classification via orthogonal transfer. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011.*, pages 801–808. Omnipress.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, 68(1):49–67.