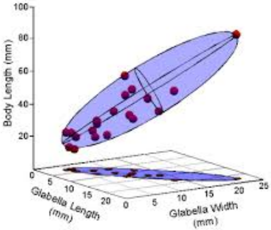
Robust PCA via Lagrange duality

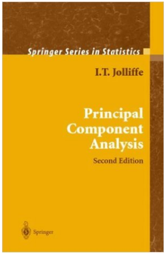
Stéphane Chrétien and Tianwen Wei
National Physical Laboratory and *Wuhan University*

Plan of the talk

PCA

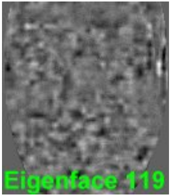
- Principal Components Analysis is a very widely used technique for dimension reduction in data analysis and visualization, machine learning, signal processing, etc ...





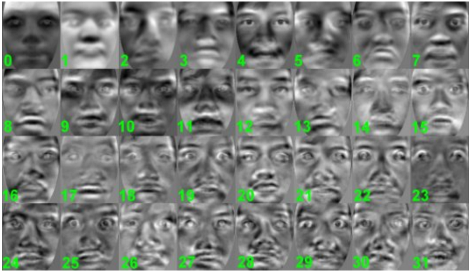
PCA

- An application example: Eigenface for face recognition



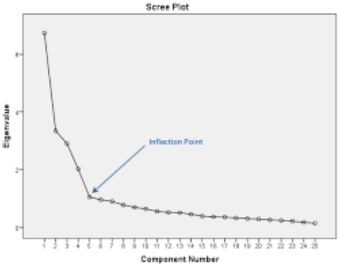
- In these example images above you can see the average face and the first and last eigenfaces that were generated from a collection of 30 images each of 4 people.
- Notice that the average face will show the smooth face structure of a generic person, the first few eigenfaces will show some dominant features of faces, and the last eigenfaces (eg: Eigenface 119) are mainly image noise.

PCA



- Eigenfaces figures out the main differences between all the images in the training set,
- One can then efficiently represent each training image using a combination of those differences (eigen-images).

PCA



- One neat property is that most eigen-faces are noise-like and only a few are associated with high eigenvalues of the covariance matrix
- Given a new image, one can project it onto the space generated by the most relevant eigenfaces and find the closest projection of an image in the training set to identify the face !

PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- The SVD is another view point on PCA
- A very fast method based on a randomized algorithm has been proposed by Candès and Witten (many previous contributions in the litterature)

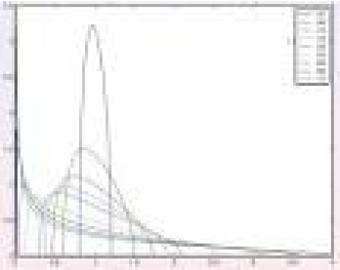
Algorithmica (2015) 72:264–281
DOI 10.1007/s00453-014-9891-7

Randomized Algorithms for Low-Rank Matrix Factorizations: Sharp Performance Bounds

Rafi Witten · Emmanuel Candès

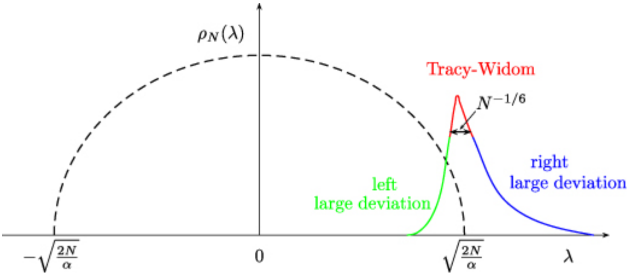
PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- For Gaussian iid data, the eigenvalues induce a random empirical measure which has been studied extensively
- The empirical spectrum converges to the Marchenko Pastur law (even universality under the fourth moment condition by Tao, Vu and Wang)



PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- For Gaussian iid data, the eigenvalues induce a random empirical measure which has been studied extensively
- The limit distribution of the maximal eigenvalue is the Tracy-Widom distribution (Johnstone, El Karoui, . . .)



PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- For Gaussian iid data, the eigenvalues induce a random empirical measure which has been studied extensively
- The spacings between the successive zeros of the Laguerre polynomials *(defining an empirical distribution converging to the Marchenko-Pastur law)* have been studied with Sébastien Darses

PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 143, Number 10, October 2015, Pages 4383–4388
http://dx.doi.org/10.1090/proc/12574
Article electronically published on April 21, 2015

ON THE SPACINGS BETWEEN THE SUCCESSIVE ZEROS OF THE LAGUERRE POLYNOMIALS

STÉPHANE CHRÉTIEN AND SÉBASTIEN DARSES

(Communicated by Sergei K. Suslov)

Abstract. We propose a simple uniform lower bound on the spacings between

PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- The Spiked model provides a very interesting example of a phase transition phenomenon *(discovered by Baik, Ben Arous and Peche)* on the detectability of a signal into a noisy environment

Theorem

Assume $X_i = U_i v + \sigma \epsilon_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ *and* $d/n = c$. *Then, as* $n \to \infty$,

$$\lambda_{\max}(Covariance) \to \begin{cases} \sigma^2(1+\sqrt{c})^2 & \text{if } \|v\|_2 \le \sigma c^{1/4} \\ (\|v\|_2 + \sigma^2)(1 + c\sigma^2/\|v\|_2^2) & \text{otherwise} \end{cases}$$

PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- PCA has also been modified in order to impose that the principal components be combinations of just a few original coordinates (variables) and the resulting method is called Sparse PCA.
- Semi-Definite Programming relaxations have been proposed by D'Aspremont, El Ghaoui, Jordan and Lanckriet

A DIRECT FORMULATION FOR SPARSE PCA USING SEMIDEFINITE PROGRAMMING*

ALEXANDRE D'ASPREMONT†, LAURENT EL GHAOUI‡, MICHAEL I. JORDAN§, AND GERT R. G. LANCKRIET¶

Abstract. Given a covariance matrix, we consider the problem of maximizing the variance explained by a particular linear combination of the input variables while constraining the number of nonzero coefficients in this combination. This problem arises in the decomposition of a covariance matrix into sparse factors or sparse PCA, and has wide applications ranging from biology to finance. We use a modification of the classical variational representation of the largest eigenvalue of a symmetric matrix, where cardinality is constrained, and derive a semidefinite programming based relaxation for our problem. We also discuss Nesterov's smooth minimization technique applied to the semidefinite program arising in the semidefinite relaxation of the sparse PCA problem. The method has complexity $O(n^4\sqrt{\log(n)}/\epsilon)$, where n is the size of the underlying covariance matrix, and ϵ is

PCA

- PCA is also a fascinating topic from both the algorithmic and theoretical perspective !
- See also the paper by Berthet and Rigollet on these topics

The Annals of Statistics
2013, Vol. 41, No. 4, 1780–1815
DOI: 10.1214/13-AOS1127

OPTIMAL DETECTION OF SPARSE PRINCIPAL COMPONENTS IN HIGH DIMENSION

BY QUENTIN BERTHET[1] AND PHILIPPE RIGOLLET[2]

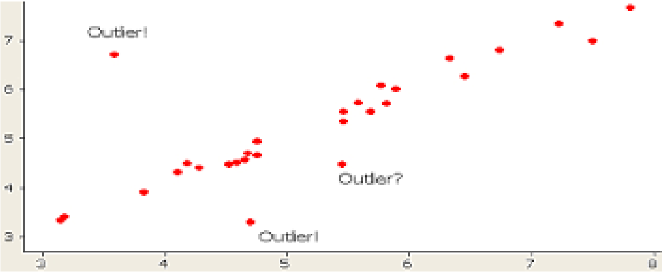
Princeton University

We perform a finite sample analysis of the detection levels for sparse principal components of a high-dimensional covariance matrix. Our minimax optimal test is based on a sparse eigenvalue statistic. Alas, computing this test is known to be NP-complete in general, and we describe a computationally efficient alternative test using convex relaxations. Our relaxation is also proved to detect sparse principal components at near optimal detection levels, and it performs well on simulated datasets. Moreover, using polynomial time reductions from theoretical computer science, we bring significant evidence that our results cannot be improved, thus revealing an inherent trade off between statistical and computational performance.

1. Introduction. The sparsity assumption has become preponderant in modern, high-dimensional statistics. In the high dimension, low sample size setting, where consistency seems to be hopeless, sparsity turns out to be the statistician's

ROBUST PCA

- However, it is known to be very sensitive to perturbations, e.g. Outliers.

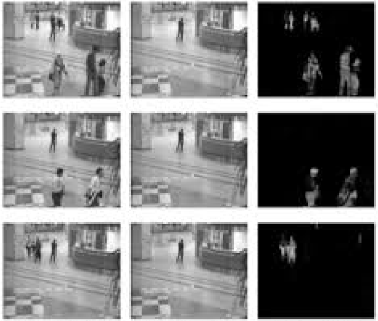


Robust PCA

- In a 2010 paper, Candès, Li, Ma, Wright studied a new version of PCA, called Robust PCA where the question of efficiently removing outliers is approached via convex optimization.

Robust PCA

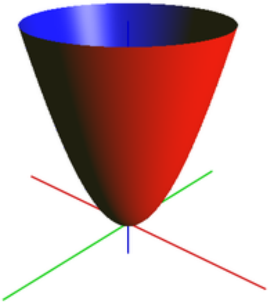
- Quick detection of outliers in high dimensions is very important at the National Physical Laboratory.
- Many researchers have worked on various sexy applications



(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$

Robust PCA

- The magic behind Robust PCA is just convex optimization



Robust PCA

- In the original version of Candès et al., the problem is the one of Low Rand + Sparse decomposition

$$\min_{L,S} \quad \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad X = L + S$$

- the nuclear norm is

$$\|L\|_* = \sum_{k=1}^{rank(L)} \sigma_k(L),$$

- the $\|\cdot\|_1$-norm is

$$\|S\|_1 = \sum_{i,j=1}^{n,m} |S_{i,j}|.$$

ROBUST PCA

- Why these norms ?
  - The norm ℓ_1 on $\mathbb{R}^n$ is the convex envelope of the *cardinal of the support* on the ℓ_∞-ball.
  - Thus, the nuclear norm is the best approximation in a certain sense of the cardinal of the support of the singular spectrum
  - ... and the ℓ_1 is the best approximation in a certain sense of the cardinal of the support of the set of matrix entrees.
- Why an exact decomposition ?
  - Can we add some potential noise ?
  - In this case, we can try and solve

$$\min_{L,S} \quad \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad \|X - L - S\|_F \leq \eta.$$

Robust PCA

- Let $L_0 = U\Sigma V^t$ be the SVD of L_0 and r its rank. Suppose that $X = L_0 + S_0$ where L_0 is $d \times n$, obeys

$$\max_i \|U^t e_i\|^2 \leq \frac{\mu r}{d} \qquad \max_i \|V^t e_i\|^2 \leq \frac{\mu r}{n}$$

$$\|UV^t\|_\infty \leq \sqrt{\frac{\mu r}{dn}}$$

and that the support set of S_0 is uniformly distributed among all sets of cardinality s.

Theorem (Candes et al. (2010))

With probability at least $1 - cn^{-10}$ (over the choice of support of S_0), Robust PCA with $\lambda = 1/\sqrt{\max\{d, n\}}$ is exact, provided that

$$\operatorname{rank}(L0) \leq \rho_r \min\{d, n\} \mu^{-1} (log \max\{d, n\})^2.$$

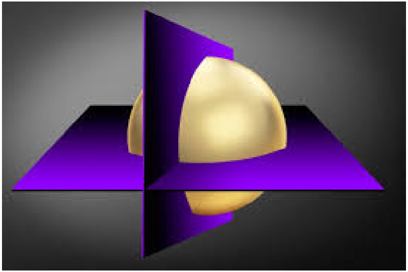
and $s \leq \rho_s dn$.

GOAL OF THE PRESENT WORK

- The proof of the Theorem of Candès et al. is quite intricate although it relies on the Golfing scheme of Gross
- Since the original paper, different approaches have emerged for other problems involving sparsity, e.g. the descent cone/gaussian mean width approach of Amelunxen et al.
- In the present work, we suggest a simple analysis of Robust PCA in the noisy setting based on the convex geometric setting of Amelunxen et al.

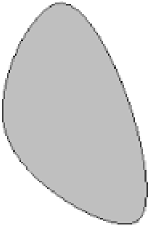
Goal of the present work

- The approach will also use Lagrange duality and a nice formula for the infimum of quadratic functionals over the sphere due to Hager.

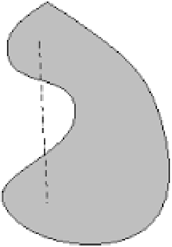


SOME CONVEX ANALYSIS

- Convexity



a convex region



a non-convex region

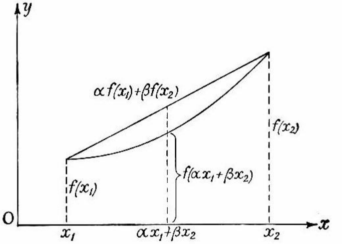
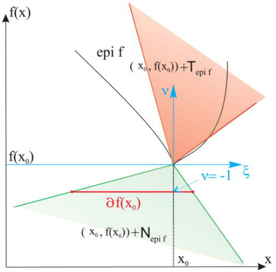


Fig. 19.—A function which is convex downwards

Some convex analysis

- Subdifferential

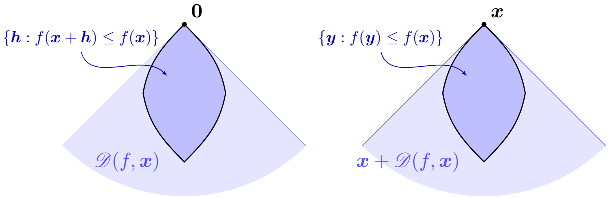
$$\partial f(x) = \left\{ g \mid f(y) \geq f(x) + \langle g, y - x \rangle \right\}$$



Some convex analysis

- Descent cone

$$\mathcal{D}f(x) = \left\{ h \mid f(x + \epsilon h) \leq f(x) \text{ for some } \epsilon > 0 \right\}$$



Descent cones and polarity

Definition (Polarity)

Given a cone K in a euclidean space $\mathbb{E}$, the polar cone K° is given by

$$K^\circ = \Big\{ y \mid \langle y, x \rangle \leq 0 \quad \forall x \in K \Big\}.$$

Proposition (Descent cone and subdifferential)

We have

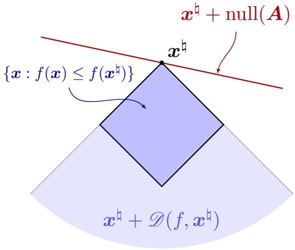
$$\mathcal{D}(f, x) = \overline{\mathrm{cone}}(\partial f(x))^\circ.$$

The descent cone approach to recovery

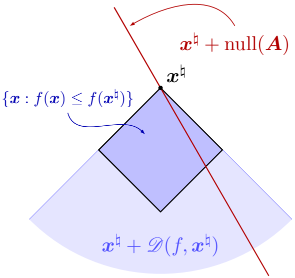
- Consider the inverse problem $y = Ax^\sharp + z$
- the estimator is taken as

$$\hat{x} = \min_x f(x) \quad \text{s.t.} \quad \|y - Ax\| \leq \eta.$$

- in the case $\eta = 0$



Success!



Failure!

The descent cone approach to recovery

- Consider the inverse problem $y = Ax^{\sharp} + z$
- the estimator is taken as

$$\hat{x} = \min_x f(x) \quad \text{s.t.} \quad \|y - Ax\| \le \eta.$$

- We will also need the following definition for the smallest eigenvalue $\lambda_{\min}(\Phi, K)$ of a linear map Φ with respect to a cone K.

$$\lambda_{\min}(\Phi, K) = \min_{\substack{\|x\|=1 \\ x \in K}} \quad \|\Phi(x)\|_2. \tag{1}$$

THE DESCENT CONE APPROACH TO RECOVERY

- Consider the inverse problem $y = Ax^\sharp + z$
- the estimator is taken as

$$\hat{x} = \min_x f(x) \quad \text{s.t.} \quad \|y - Ax\| \leq \eta.$$

- A recent result of Tropp et al. is the following theorem:

Theorem (Tropp et al.)

Assume that $y = \Phi(x^\sharp) + z$ and that $\|z\|_2 \leq \eta$. Let

$$\hat{x} \in \operatorname{argmin}_{x \in \mathbb{E}} f(x) \quad s.t. \quad \|y - \Phi(x)\|_2 \leq \eta.$$

Then, we have

$$\|\hat{x} - x^\sharp\|_F^2 \leq \frac{2\eta}{\lambda_{\min}(\Phi, \mathcal{D}(f, x_0))}.$$

A lower bound on $\lambda_{\min}(\Phi, K)$

- For Robust PCA, the observation operator is very simple:

$$\Phi(L, S) = L + S$$

- A main difference with previous work based on the descent cone and the conic singular value is that

Φ is not random !

A lower bound on $\lambda_{\min}(\Phi, K)$

- The sparsity promoting penalty is

$$f(L, S) = \|L\|_* + \lambda \|S\|_1.$$

- We have

$$\lambda_{\min}\Big(\Phi, \mathcal{D}(f, (L_0, S_0))\Big)^2 = \min_{\substack{\|L\|_F^2 + \|S\|_F^2 = 1; \\ (L,S) \in \mathcal{D}(f, (L_0, S_0))}} \|\Phi(L, S)\|_F^2$$

$$= \min_{\substack{\|L\|_F^2 + \|S\|_F^2 = 1; \\ (L,S) \in \mathcal{D}(f, (L_0, S_0))}} \|L + S\|_F^2$$

A lower bound on $\lambda_{\min}(\Phi, K)$

- By the rules of subdifferential calculus, we have

$$\partial f(L_0, S_0) = \partial\|\cdot\|_*(L_0) \times \lambda\, \partial\|\cdot\|_1(S_0)$$

  with

  - $$\partial\|\cdot\|_*(L_0) = \{U_0 V_0^t + W_0 \mid \|W_0\| \leq 1,\ U_0^t W = 0,\ W_0 V_0 = 0\}$$

    where $L_0 = U_0 \Sigma_0 V_0^t$ is a SVD of L_0,
  - and

$$\partial\|\cdot\|_1(S_0) = \mathrm{sign}(S_0).$$

A lower bound on $\lambda_{\min}(\Phi, K)$

- Consider the Lagrange function:

$$(L, S, \Gamma_L, \Gamma_S) = \|L + S\|_F^2 + \langle \Gamma_L, L \rangle + \langle \Gamma_S, S \rangle$$

subject to the implicit constraint

$$\|L\|_F^2 + \|S\|_F^2 = 1$$

- and the associated dual function

$$\Theta(\Gamma_L, \Gamma_S) = \inf_{\substack{L,\, S \in \mathbb{R}^{d \times n} \\ \|L\|_F^2 + \|S\|_F^2 = 1}} \mathcal{L}(L, S, \Gamma_L, \Gamma_S).$$

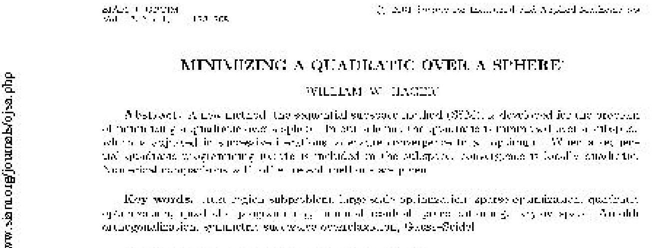
- Moreover, the dual variables Γ_L and Γ_S should be constrained to lie in the polar cone $\mathcal{D}(f, (L_0, S_0))^\circ$.

A lower bound on $\lambda_{\min}(\Phi, K)$

- Although presented as a nonconvex problem, the infimum in

$$\Theta(\Gamma_L, \Gamma_S) = \inf_{\substack{L,\, S \in \mathbb{R}^{d \times n} \\ \|L\|_F^2 + \|S\|_F^2 = 1}} \mathcal{L}(L, S, \Gamma_L, \Gamma_S).$$

  has in fact a closed form expression
- this is well described in a previous work on spherically



MINIMIZING A QUADRATIC OVER A SPHERE

WILLIAM W. HAGER

A lower bound on $\lambda_{\min}(\Phi, K)$

- Let $\gamma_L = \mathrm{vec}(\Gamma_L)$, $\gamma_C = \mathrm{vec}(\Gamma_C)$ in $\mathbb{R}^{dn}$.
- Thus, the dual function Θ is given by

$$\Theta(\Gamma_L, \Gamma_S) = \inf_{\substack{z \in \mathbb{R}^{2dn} \\ z^t z = 1}} z^t \mathfrak{Q} z - 2\gamma^t z$$

  - with

$$\mathfrak{Q} = \begin{bmatrix} I \\ I \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix}^t$$

  - and

$$\gamma = -\frac{1}{2} \begin{bmatrix} \gamma_L \\ \gamma_C \end{bmatrix}.$$

A lower bound on $\lambda_{\min}(\Phi, K)$

- The matrix $\mathfrak{Q}$ has the eigenvalue decomposition

$$\mathfrak{Q} = U \Lambda U^t,$$

with

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} I & I \\ I & -I \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} 2I & 0 \\ 0 & 0 \end{bmatrix}$$

A lower bound on $\lambda_{\min}(\Phi, K)$

- Following Hager's paper, if $\gamma_L - \gamma_C = 0$, the solution is given by

$$z^* = U \begin{bmatrix} c_L^* \\ c_S^* \end{bmatrix} \quad \text{with} \quad \begin{cases} c_L^* = -\frac{1}{\sqrt{2}} \gamma_L \\ \\ c_S^* \end{cases},$$

and

$$\frac{\|\gamma_L\|_2^2}{2} + \|c_S^*\|_2^2 = 1. \qquad (2)$$

- Another expression for z^* is

$$z^* = U \begin{bmatrix} \frac{1}{\sqrt{2}} \gamma_L \\ c_S^* \end{bmatrix}$$

A LOWER BOUND ON $\lambda_{\min}(\Phi, K)$

- Using this formula, one obtains that

$$\Theta(\Gamma_L, \Gamma_S) = 2\,\|\Gamma_L\|_F^2.$$

- Notice also that we should satisfy the following constraints

$$\Gamma_L = \Gamma_S$$

$$\frac{\|\Gamma_L\|_2^2}{2} + \|c_S^*\|_2^2 = 1.$$

  when this problem is feasible.
- Therefore, if the above constraints can be satisfied, the optimal value of **the dual function Θ has optimal value which will simply be equal to 4**.

When do we have $\Gamma_L = \Gamma_S$?

- We now address the question of feasibility of the constraints

$$\Gamma_L = \Gamma_S$$

- Recall that we should have the constraint that

$$(\Gamma_L, \Gamma_S) \in \overline{\text{cone}}(\partial f(L_0, S_0)).$$

WHEN DO WE HAVE $\Gamma_L = \Gamma_S$?

- Moreover, we have

$$\overline{\text{cone}}(\partial f(L_0, S_0)) = \Big\{ \mu\,(\Gamma_L, \Gamma_S) \mid \Gamma_L \in \Big\{ U_0 V_0^t + W_0 \mid U_0^t W_0 = 0,\ W_0 V_0 = 0,\ \|W_0\| \leq 1 \Big\}, \; \Gamma_S \in \lambda\,\text{sign}(S_0) \text{ and } \mu \in \mathbb{R}_+ \Big\}.$$

When do we have $\Gamma_L = \Gamma_S$?

- We can take W_0 as a solution of the following system

$$\left(U_0 V_0^t + W\right)_{\Omega} = \lambda\, \mathrm{sign}(S_0)_{\Omega}, \quad |\left(U_0 V_0^t + W\right)_{\Omega^c}| \le \lambda$$

  and

$$U_0^t W = 0,\ W V_0 = 0, \tag{3}$$

  where Ω is the support of S_0.

- The constraints (3) are equivalent to

$$W = U_0^{\perp} C {V_0^{\perp}}^t.$$

WHEN DO WE HAVE $\Gamma_L = \Gamma_S$?

- Let us first consider a toy case, namely when $U_0 V_0^t = 0$.
- In this case, we look for a solution C to the feasibility problem

$$-\lambda \leq \left(U_0^{\perp} C V_0^{\perp^t} \right)_{i,i'} \leq \lambda, \quad (i,i') \in \Omega^c \tag{4}$$

and

$$\left(U_0^{\perp} C V_0^{\perp^t} \right)_{i,i'} = \lambda \operatorname{sign}(S_0)_{i,i'}, \quad (i,i') \in \Omega. \tag{5}$$

When do we have $\Gamma_L = \Gamma_S$?

- Notice that we can rewrite the system given by (4) and (5) as

$$\Omega^c\left(U_0^{\perp} C V_0^{\perp^t}\right) \leq \lambda e, \tag{6}$$

$$-\Omega^c\left(U_0^{\perp} C V_0^{\perp^t}\right) \leq \lambda e \tag{7}$$

and

$$\Omega\left(U_0^{\perp} C V_0^{\perp^t}\right) = \lambda\, \Omega\Big(\mathrm{sign}(S_0)\Big). \tag{8}$$

WHEN DO WE HAVE $\Gamma_L = \Gamma_S$?

- Let us also define the operator $\mathcal{A} : \mathbb{R}^{r_0 \times r_0} \mapsto \mathbb{R}^{d \times n}$ as given by

$$\mathcal{A}(C) = U_0^{\perp} C V_0^{\perp^t}. \tag{9}$$

Its adjoint is easily seen to be given by

$$\mathcal{A}^*(X) = U_0^{\perp^t} X V_0^{\perp}. \tag{10}$$

WHEN DO WE HAVE $\Gamma_L = \Gamma_S$?

- Taking the dual approach, we may now use the Farkas lemma to study the feasibility of (4) and (5).

Theorem 2.4 (Farkas Lemma)[2] *Let A be an $m \times n$ matrix, $b \in \mathbb{R}^m$ and $F = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$. Then either $F \neq \emptyset$ or there exists $y \in \mathbb{R}^m$ such that $yA \geq 0$ and $y \cdot b < 0$ but not both.*

Proof First we prove that both statements cannot hold simultaneously. Suppose not. Let $x^* \geq 0$ be a solution to $Ax = b$ and y^* a solution to $yA \geq 0$ such that $y^* b < 0$. Notice that x^* must be a solution to $y^* Ax = y^* b$. Thus $y^* Ax^* = y^* b$. Then $0 \leq y^* Ax^* = y^* b < 0$, a contradiction.

- Thus (4) and (5) will be infeasible if there exist z_+, z_- in $\mathbb{R}_+^{dn-s_0}$ and y in $\mathbb{R}^{s_0}$ such that

$$\mathcal{A}^* \circ \Omega^{c^*}(z_1 - z_2) + \mathcal{A}^* \circ \Omega^*(y) = 0 \tag{11}$$

and

$$e^t(z_1 + z_2) + \Omega(\mathrm{sign}(S_0))^t y < 0. \tag{12}$$

WHEN DO WE HAVE $\Gamma_L = \Gamma_S$?

- Moreover, infeasibility of (4) and (5) is equivalent to the existence of z in $\mathbb{R}^{dn-s_0}$ and y in $\mathbb{R}^{s_0}$ such that

$$U_0^{\perp^t} \mathcal{X}(z, y) V_0^{\perp} = 0 \tag{13}$$

and

$$\|z\|_1 + \Omega(\mathrm{sign}(S_0))^t y < 0. \tag{14}$$

When do we have $\Gamma_L = \Gamma_S$?

- To handle this expression easily, one may want to use a Null Space-type condition

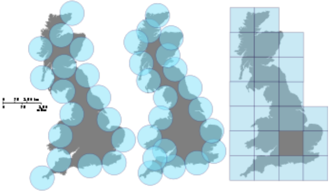
$$U_0^{\perp t} C V_0^{\perp} = 0 \quad \Rightarrow \quad \|C_\Omega\|_2 \leq \frac{\rho}{\sqrt{s_0}} \|C_{\Omega^c}\|_1 \tag{15}$$

- Thus, we obtain that $\|y\|_2 \leq \frac{\rho}{\sqrt{s_0}} \|z\|_1$ and using Hoeffding's inequality, we get

$$\mathbb{P}\left(\frac{\|z\|_1 + \Omega(\operatorname{sign}(S_0))^t y}{\|z\|_1} \leq \left(1 - \rho \sqrt{2 \frac{\log(1/\alpha)}{s_0}} \right) \right) \leq \alpha.$$

When do we have $\Gamma_L = \Gamma_S$?

- Now, we have to consider the infimum of $1 + \Omega(\mathrm{sign}(S_0))^t y / \|z\|_1$ over all possible values of z and y satisfying the condition $\|y\|_2 \leq \rho / \sqrt{s_0}\ \|z\|_1$.
- Due to the Hoeffding bound, one only needs to consider the infimun of $1 + \Omega(\mathrm{sign}(S_0))^t \tilde{y}$ for all $\tilde{y}$ in the euclidean ball $B_2(\rho/\sqrt{s_0})$.
- For this purpose, we will simply introduce an ϵ-net and use the union bound.



When do we have $\Gamma_L = \Gamma_S$?

- For any $\epsilon > 0$, it is known that there exists an ϵ-net of size at most $(3/\epsilon)^{s_0}$ such that the ball $B_2(1)$ is covered by balls of radius ϵ centered at the points of the net.
- By dilation, we deduce that we only need at most the same number of points to obtain a $\rho\epsilon/\sqrt{s_0}$-net, denoted by $\mathcal{N}(\rho\epsilon/\sqrt{s_0})$, which may be used to cover the ball $B_2(\rho/\sqrt{s_0})$.
- Using the union bound, we get

$$\mathbb{P}\left(\min_{\tilde{y}\in\mathcal{N}(\rho\epsilon)} 1 + \Omega(\mathrm{sign}(S_0))^t\tilde{y} \leq \left(1 - \rho\sqrt{2\,\frac{\log(1/\alpha)}{s_0}}\right)\right)$$

$$\leq \alpha\,(3/\epsilon)^{s_0}.$$

When do we have $\Gamma_L = \Gamma_S$?

- Moreover, since every point in the ball $B_2(\rho/\sqrt{s_0})$ is at most at a distance of $\rho\epsilon/\sqrt{s_0}$ from a point in the net $\mathcal{N}(\rho\epsilon/\sqrt{s_0})$, using the Cauchy-Schwartz inequality, we get

$$\max_{\tilde{y}\in B_2(\rho/\sqrt{s_0})} \min_{\tilde{\tilde{y}}\in \mathcal{N}(\rho\epsilon/\sqrt{s_0})} \left|\Omega(\mathrm{sign}(S_0))^t\tilde{y} - \Omega(\mathrm{sign}(S_0))^t\tilde{\tilde{y}}\right| \leq \rho\epsilon.$$

- From this, we obtain

$$\mathbb{P}\left(\min_{\tilde{y}\in B_2(\rho/\sqrt{s_0})} 1+\Omega(\mathrm{sign}(S_0))^t\tilde{y} \leq 1-\rho\sqrt{2\,\frac{\log(1/\alpha)}{s_0}}-\rho\epsilon\right) \leq \alpha\,(3/\epsilon)^{s_0}.$$

- Therefore, we deduce that if $\rho < \sqrt{\epsilon/4}$ the system is infeasible, with probability at most $(\epsilon/3)^{s_0}$.

When do we have $\Gamma_L = \Gamma_S$?

- In the effective case (not the toy model), assuming that

$$-\frac{\lambda}{2} \leq \Omega^c(U_0 V_0) \leq \frac{\lambda}{2}. \tag{16}$$

- and doing the same computations again, we obtain

$$\mathbb{P}\left(\min_{\tilde{y} \in B_2(\rho/\sqrt{s_0})} 1 + \Omega(\mathrm{sign}(S_0))^t \tilde{y} \leq \frac{1}{2} - \rho \sqrt{\frac{2\ \log(1/\alpha)}{s_0}} - \rho\epsilon \right) \leq \alpha\ (3/\epsilon)^{s_0}. \tag{17}$$

- Therefore, we deduce from this computation that if $\rho < \sqrt{\epsilon/16}$ the system is infeasible, with probability at most $(\epsilon/3)^{s_0}$.

Summing up !

- What did we need ?
  - $-\frac{\lambda}{2} \leq \Omega^c(U_0 V_0) \leq \frac{\lambda}{2}$
  - The (ρ, s_0)-Null Space property of $C \mapsto {U_0^{\perp}}^t C V_0^{\perp}$ with $\rho < \sqrt{\epsilon/16}$
- What did we obtain ?

$$\|(\hat{L}, \hat{S}) - (L_0, S_0)\|_F^2 \leq \frac{2\eta}{4}.$$

with probability at most $(\epsilon/3)^{s_0}$

What about the Null-Space condition ?

- Note that the Null Space condition is equivalent to

$$\forall (i,j),\ \operatorname{trace}\left(C V_0^{\perp} E_{i,j} U_0^{\perp^t}\right) = 0 \quad \Rightarrow \quad \|C_\Omega\|_2 \leq \frac{\rho}{\sqrt{s_0}} \|C_{\Omega^c}\|_1.$$

- Let M denote the matrix whose rows are the row vectors. $\operatorname{vec}(U_0^{\perp} E_{j,i} V_0^{\perp^t})^t$
- Let $c = \operatorname{vec}(C)$.
- Let ω denote the index set of the components of c satisfying $c_\omega = C_\Omega$.
- Then, the Null Space condition can be rewritten as

$$Mc = 0 \quad \Rightarrow \quad \|c_\omega\|_2 \leq \frac{\rho}{\sqrt{s_0}} \|c_{\omega^c}\|_1.$$

WHAT ABOUT THE NULL-SPACE CONDITION ?

- Moreover, we have

$$U_0^{\perp} E_{j,i} V_0^{\perp^t} = U_{0,j}^{\perp} V_{0,i}^{\perp^t}.$$

where $U_{0,j}^{\perp}$ is the j^{th} column of $U_0^{\perp}$ and $V_{0,i}^{\perp}$ is the i^{th} column of $V_{0,i}$.

- Therefore, the row of M indexed by (i,j), further denoted by $m_{i,j}$ is the row vector whose components are the product of the components of $U_{0,j}$ with the components of $V_{0,i}$.

What about the Null-Space condition ?

- The matrix M has orthogonal rows. Indeed, let us compute the scalar product of the row indexed by (i,j) with the row indexed by (i',j')

$$\begin{aligned}\langle m_{i,j}, m_{i',j'}\rangle &= \sum_{k=1}^{r}\sum_{k'=1}^{r} U_{0,j,k}V_{0,i,k'}U_{0,j',k}V_{0,i',k'} \\ &= \left(\sum_{k=1}^{r} U_{0,j,k}U_{0,j',k}\right)\left(\sum_{k'=1}^{r} V_{0,i,k'}V_{0,i',k'}\right)\end{aligned}$$

- Due to the orthogonality of the columns of U_0 and V_0,

$$\langle m_{i,j}, m_{i',j'}\rangle = 0$$

What about the Null-Space condition ?

- The number of rows of $(M_{ij})_{(i,j)\in\Omega^c}$ is $(\min\{d,n\}-r_0)^2-s_0$ and it has dn columns.
- Since its rows form an orthonormal matrix in $\mathbb{R}^{dn\times dn}$, and after rescaling, an orthogonal matrix satisfies $2s_0$-Restricted Invertibility Property for any s_0 less than $dn/2$ and
- If r_0 and s_0 are small compared to $\min\{d,n\}$, the matrix $(M_{ij})_{(i,j)\in\Omega^c}$ has only a few less rows than an orthogonal matrix, it seems reasonable to require the $2s_0$-Restricted Invertibility Property for some s_0 in the applications.
- Since the $2s_0$-Restricted Invertibility Property implies the s_0-Null Space Property, this Null Space Property seems reasonable in the application as well.

CONCLUSIONS ?

- PCA is still full of surprises and is fun to analyse !
- We obtained a result which depends on a Null Space Property of the original eigenvectors: plug in your model of choice and obtain the possible values of r_0 and s_0 !
- If we use delocalization types of results (eigenvectors have components with same magnitude), we obtain
  - that λ is in the range of values obtained in the paper by Candès, Li, Ma, Wright.
  - the Null space property holds for $s_0 \leq C_s(\rho) \min\{d, n\}^2$ and $r_0 \leq C_r(\rho) \min\{d, n\}/\log(\min\{d, n\})$, which is of the same order as in the paper by Candès, Li, Ma, Wright (up to a log !), using a modification of a famous theorem of Rudelson and Vershynin on RIP for Discrete Fourier Transform matrices.
- but ... is delocalization relevant for all applications ?

Conclusions ?

- PCA is still full of surprises and is fun to analyse !
- We obtained a result which depends on a Null Space Property of the original eigenvectors: plug in your model of choice and obtain the possible values of r_0 and s_0 !
- The Null Space condition could be addressed by other means than RIP and (?) because we only need such a property to hold for the support of the outliers *(RIP is in fact much stronger than what we really need !)*
- The main goal of this work is to provide a simple method for addressing low-rand + more complex sparse structured models for which we don't know how to extend the approach by Candès et al.

Conclusions ?

Thank you for your attention !