INITIAL RESEARCH PROJECT REPORT

Exact Recovery for Binary Spiked Wishart Model

Author:
MARIN BALLU

Supervisor:
QUENTIN BERTHET

Department of Pure Mathematics and Mathematical Statistics
Cambridge Center for Analysis
Cantab Capital Institute for Mathematics of Information

January 2019
Cambridge, UK

Contents

Overview

This report is a review of the first problem that I studied during my first year of PhD at the University of Cambridge. It is about recovering the largest eigenvector of a covariance matrix, under the assumption that this vector has binary entries. This is a variant of the already solved case of the largest eigenvector being sparse, and can be interpreted as a meaningful problem in population detection. As the problem is NP-complete, we relax it using semi-definite programming in order to compute a solution in polynomial time. We proved statistical optimality of such a solution, and provided an algorithm to obtain it.

1 Introduction

Abstract. We consider the problem associated to recovering the covariance matrix of i.i.d. Gaussian vectors such that all but one eigenvalue of this matrix is one, and that the eigenvector associated with the non trivial

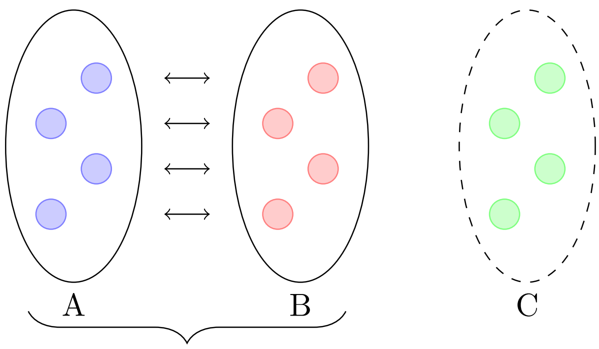
eigenvalue in in the binary hypercube. Our method involves a SDP relaxation, and we will proceed on giving a statistical upper bound and lower bound for our estimator, as well as a computational upper bound.

Context. This follows the work of Berthet, Rigollet and Srivastava [2] on the similar problem of recovering the block structure of an Ising blockmodel given independent observations of the binary cube. A general analysis of spiked Wishart distribution has been provided by Mo [10]. Put aside the probabilistic aspects, the SDP method was also used for the broader scope of quadratic minimization over the integer lattice ^{p}, in a recent paper from Park and Boyd [11]. On a side note, our problem can be translated as a MAXCUT problem as solved by Goemans [7].

Notations. Let $p \in$ be the dimension of our problem, $n \in$ the sample size, $v \in \{-1, 1\}^p$ a binary vector and $\lambda > 0$ the scale parameter. Let $X_1, \ldots, X_n$ be i.i.d. vectors of ^{p} following the $p-$variate normal distribution $\mathcal{N}_p(0, \Sigma)$. Then the *Wishart distribution with degree of liberty n and with covariance matrix* Σ is the probability distribution $W_p(n, \Sigma)$ of the $p \times p$ random matrix $\frac{1}{n}\sum_{k=1}^n X_k X_k^T$ [14]. It is the law of the sample covariance matrix. The *spiked Wishart distribution, with spike v and intensity λ,* is $W_p(n, \Sigma_v)$ with $\Sigma_v = I_p + \frac{\lambda}{p} v v^T$ the spiked covariance matrix.

From now on let $X_1, \ldots, X_n \sim \mathcal{N}_p(0, \Sigma_v)$ be i.i.d. vectors, and $\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^n X_k X_k^T$ the sample covariance matrix. We show a method to recover exactly $\pm v$ with high probability.

Interpretation. We can interpret this problem as a population detection problem.



We have p individuals, with an activity $X^{(1)}, \ldots, X^{(p)}$ following a one dimensional normal distribution $\mathcal{N}(0, 1 + \lambda)$. There exists two groups of individuals A and B in competition so that if $i \neq j$ are two individuals in the same community $\mathbb{E}\left[X^{(i)} X^{(j)}\right] = \lambda$ and if they are in an opposite community $\mathbb{E}\left[X^{(i)} X^{(j)}\right] = -\lambda$. The goal is to recover exactly the two groups of individuals A and B, with high probability, if we sample the same system n independent times. The vector v represents the two groups with $v_i = 1$ if $i \in A$ and $v_i = -1$ if $i \in B$, and the distribution of the random vector $X = (X^{(1)}, \dots, X^{(p)})$ is thus our sample distribution $\mathcal{N}_p(0, \Sigma_v)$.

As an option, we can add a third group C so that the activity of any individual from C is independent from the rest of the individuals in $A \cup B$. We can first select $A \cup B$ with a sparse PCA [1], thus the problem remains the same.

2 Estimator

2.1 General strategy

Naturally one can consider

$$\hat{v} = \arg \max_{u \in \{-1,1\}^p} u^T \hat{\Sigma} u,$$

and then prove $\hat{v} = v$ with high probability. The maximization of a non specified quadratic form over the set of binary vectors is known to be computationally NP-hard (see MAX CUT [7]). However this problem can be relaxed as a convex minimization problem over a convex set. We note that $u^T \hat{\Sigma} u = \mathrm{tr}(\hat{\Sigma} u u^T)$, so

$$\hat{V} = \arg \max_{U \in \mathcal{E}} \mathrm{tr}\, \hat{\Sigma} U,$$

with $\hat{V} = \hat{v}\hat{v}^T$ and $\mathcal{E} = \{U = uu^T : u \in \{-1,1\}^p\}$. As $\hat{v} \in \{-1,1\}^p$, the rows and columns of the matrix $\hat{V}$ are equal to $\hat{v}$ or $-\hat{v}$, so we can recover $\pm\hat{v}$ with $\hat{V}$. We now note $V = vv^T$. So the problem is now a maximization of the scalar product with $\hat{\Sigma}$ over the set $\mathcal{E}$. The relaxation consists in replacing $\mathcal{E}$ by a larger convex set $\mathcal{E}' \subset \mathcal{S}_p$, with $\mathcal{S}_p$ being the set of $p-$dimensional symmetric real matrices. Then we get a solution

$$\hat{V}' \in \arg \max_{U \in \mathcal{E}'} \mathrm{tr}\, \hat{\Sigma} U,$$

For every matrix $A \in \mathcal{E}'$, we can define the normal cone $\mathcal{C}_A$ to $\mathcal{E}'$ at the point A by

$$\mathcal{C}_A = \{S \in \mathcal{S}_p : \forall U \in \mathcal{E}',\ \mathrm{tr}(SU) \leqslant \mathrm{tr}(SA)\}.$$

As $\mathcal{E}'$ is convex, we see that $\mathcal{C}_A \neq \varnothing$ if and only if $A \in \partial\mathcal{E}'$. For information on normal cones and convex minimization in general see [4].

The topological interior of the normal cone is

$$\mathring{\mathcal{C}}_A = \{S \in \mathcal{S}_p : \forall U \in \mathcal{E}' \backslash \{A\},\ \mathrm{tr}(SU) < \mathrm{tr}(SA)\}.$$

Then we have the following assumptions:

- If $\hat{\Sigma} \in \mathring{\mathcal{C}}_V$ then $\hat{V}' = V$.

- Reciprocally if $\hat{V}' = V$ then $\hat{\Sigma} \in \mathcal{C}_V$.

For the first one we write $\forall U \in \mathcal{E}' \backslash \{V\},\ \operatorname{tr}(\hat{\Sigma} U) < \operatorname{tr}(\hat{\Sigma} V)$, so V is the only matrix that maximizes $\operatorname{tr}(\hat{\Sigma} U)$ over $U \in \mathcal{E}'$, so $\hat{V}' = V$. The second one follows by definition of $\hat{V}'$.

As $\hat{\Sigma}$ has a distribution over $\mathcal{S}_p$ that is absolutely continuous with respect to Lebesgue measure in $\mathcal{S}_p$ [14],

$$\mathbb{P}\left(\hat{\Sigma} \in \mathcal{C}_V\right) = \mathbb{P}\left(\hat{\Sigma} \in \mathring{\mathcal{C}}_V\right).$$

So, we need that $\mathring{\mathcal{C}}_V \neq \varnothing$, to have a chance to get exact recovery. The set of matrices $A \in \mathcal{E}'$ that have a normal cone with nonempty interior is the set of vertices $\operatorname{Vert}(\mathcal{E}')$ of $\mathcal{E}'$ (defined in [4]). This is a countable subset of the frontier $\partial \mathcal{E}'$: each vertex is associated to an open solid angle (the intersection of the interior of the normal cone with the unit sphere), which is an open subset of the sphere Θ_k so that each two of these subsets are disjointed, and the sphere is separable.

Let $\mathcal{E}'$ and $\mathcal{E}''$ be two convex sets with a common point A, if $\mathcal{E}' \subset \mathcal{E}''$ then $\mathcal{C}_A(\mathcal{E}'') \subset \mathcal{C}_A(\mathcal{E}')$. So we need to choose a convex set $\mathcal{E}'$ verifying $\mathcal{E} \subset \operatorname{Vert}(\mathcal{E}')$, as small as possible, and so that the scalar product maximization over this set is highly computable.

Naively we can choose $\mathcal{E}^{\square} = \{(a_{ij}) \in \mathcal{S}_p : \forall i, j,\ |a_{ij}| \leqslant 1,\ a_{ii} = 1\}$, over which a gradient descent is easy to compute and, as it is a polytope, $\hat{\Sigma}$ is almost surely in the interior of normal cone of a vertex $A \in \operatorname{Vert}(\mathcal{E}^{\square}) = \{(a_{ij}) \in \mathcal{S}_p : \forall i, j,\ |a_{ij}| = 1,\ a_{ii} = 1\}$. However the number of those vertices is $2^{n(n-1)/2}$, which is way more than $\operatorname{Card}(\mathcal{E}) = 2^{n-1}$, so intuitively the method is inefficient for exact recovery. This can be proven by direct calculus.

We then assume we chose $\mathcal{E}' \subset \mathcal{E}^{\square}$. This simple assumption gets us a crucial result:

Proposition 1. *For all choice of $\mathcal{E}' \subset \mathcal{E}^{\square}$, for all $V \in \mathcal{E}$, we have*

$$\Sigma_v = I_p + \frac{\lambda}{p} V \in \mathring{\mathcal{C}}_V.$$

Proof. Let $A = (a_{ij}) \in \mathcal{E}'$, then $|a_{ij}| \leqslant 1$, so we have

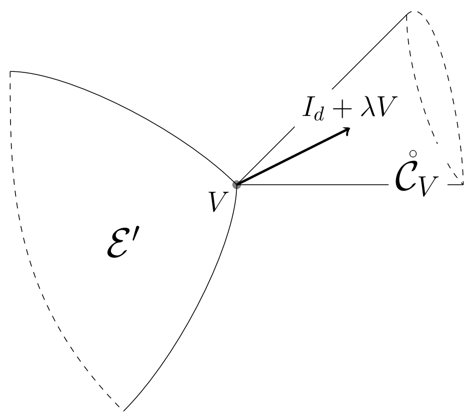
$$\operatorname{tr}(VA) = \sum_{i,j} a_{ij} v_i v_j \leqslant p^2$$

with equality if and only if $A = V$. Moreover $\operatorname{tr}(\Sigma_v V) = (1 + \lambda) p$, so

$$\operatorname{tr}(\Sigma_v A) = \operatorname{tr}(A) + \frac{\lambda}{p} \operatorname{tr}(VA) \leqslant (1 + \lambda) p = \operatorname{tr}(\Sigma_v V)$$

with equality if and only if $A = V$. So $\Sigma_v \in \mathring{\mathcal{C}}_V$. □

It graphically shows as:



We know that $\mathbb{E}\left[\hat{\Sigma}\right] = \Sigma_v$, so with concentration inequalities [3] of the type

$$\mathbb{P}\left(f\left(\hat{\Sigma} - \mathbb{E}\left[\hat{\Sigma}\right]\right) > y\right) \leqslant \delta,$$

we can guarantee that $\hat{\Sigma} \in \mathring{\mathcal{C}}_V$ with high probability. This is how we will proceed in the following sections to get a statistical upper bound.

2.2 Relaxation as a semi-definite programming problem

To choose a good relaxation space, we describe $\mathcal{E}$ in a way that we can relax without losing too much information.

Proposition 2. *For $M \in \mathcal{S}_p$, we note* $\mathrm{Diag}(M)$ *the diagonal matrix with the same diagonal entries as M.*

$$\mathcal{E} = \{U \in \mathcal{S}_p : U \succcurlyeq 0,\ \mathrm{Diag}(U) = I_p,\ \mathrm{rank}(U) = 1\}.$$

Proof. $\boxed{\subset}$ Let $V \in \mathcal{E}$, by definition of $\mathcal{E}$ we know $V = vv^T \in \mathcal{S}_p$ with $v \in \{-1, 1\}^p$. Then $\mathrm{rank}\, V = 1$, $V \succcurlyeq 0$ and $V_{ii} = v_i^2 = 1$.

$\boxed{\supset}$ As $\mathrm{rank}(U) = 1$ we can write $U = ba^T$ with a and b two $p-$dimensional vectors. As U is symmetric, the image set of U is orthogonal to the kernel set. So $b \perp (a)^{\perp}$, so $b = \epsilon a$ with $\epsilon = \pm 1$. As $U \succcurlyeq 0$ we remark $\epsilon \|a\|^2 = a^T U a \succcurlyeq 0$, so $\epsilon = 1$. We then have $a_i^2 = U_{ii} = 1$, so $a_i = \pm 1$. So finally $U \in \mathcal{E}$. □

We define the elliptope $\mathcal{E}^* = \{U \in \mathcal{S}_p : U \succcurlyeq 0,\ \mathrm{Diag}(U) = I_p\}$ [4],[5]. In this case, let's show it is a good relaxation set.

Proposition 3. *The set $\mathcal{E}^*$ verifies the following properties:*

1. *$\mathcal{E}^*$ is strictly convex,*

2. *$\mathcal{E}^*$ is a subset of $\mathcal{E}^\square$,*

3. *$\mathrm{Vert}(\mathcal{E}^*) = \mathcal{E}$.*

Proof. 1. $\mathcal{E}^*$ is the intersection of the affine space $\{U \in \mathcal{S}_p : \mathrm{Diag}(U) = I_p\}$ and the strictly convex set $\mathcal{S}_p^+$.

2. Let $U \in \mathcal{E}^*$, we already have $\mathrm{Diag}(U) = I_p$, so we just need to show $|U_{ij}| \leqslant 1$ for all $i \neq j$. Let $(e_1, \dots, e_p)$ be the canonical basis of ^{p}. U is semidefinite positive so for all $\alpha \in$, for all $i \neq j$,

$$0 \leqslant (\alpha e_i + e_j)^T U(\alpha e_i + e_j) = \alpha^2 U_{ii} + U_{jj} + 2\alpha U_{ij} = \alpha^2 + 1 + 2\alpha U_{ij},$$

we minimize over $\alpha \in$: $\alpha_{\min} = -U_{ij}$, so $1 - U_{ij}^2 \geqslant 0$, and finally $|U_{ij}| \leqslant 1$. So $U \in \mathcal{E}^\square$.

3. $\boxed{\supset}$ We know $\mathcal{E} \subset \mathcal{E}^* \subset \mathcal{E}^\square$, so for all $U \in \mathcal{E}$, $\varnothing \neq \mathcal{C}_U(\mathcal{E}^\square) \subset \mathcal{C}_U(\mathcal{E}^*)$, thus $U \in \mathrm{Vert}(\mathcal{E}^*)$.
$\boxed{\subset}$ It follows from a differential geometry argument [5] that vertices are rank 1 matrices of $\mathcal{E}^*$, which are in $\mathcal{E}$ by Proposition 2.

□

Remark. The strict convexity of $\mathcal{E}^*$ implies that $\arg\max_{U \in \mathcal{E}'} \mathrm{tr}(\hat{\Sigma}U)$ is a singleton.

So we write the final version of our problem:

$$\boxed{\hat{V}^* = \arg\max_{U \in \mathcal{E}^*} \mathrm{tr}(\hat{\Sigma}U).}$$

It is in a special class of convex optimization problems: semidefinite programming. There exist various algorithms that permit to get a good approximate solution in polynomial time (see [4], [13]).

3 Upper Bound

3.1 Analysis

We now focus on the upper bound of our estimator, bounding

$$\mathbb{P}\left(\hat{V}^* \neq V\right) = \mathbb{P}\left(\hat{\Sigma} \notin \mathring{\mathcal{C}}_V^*\right) = \mathbb{P}\left(\hat{\Sigma} \notin \mathcal{C}_V^*\right).$$

For that matter, we underline the following description of $\mathring{\mathcal{C}}_V^*$:

Proposition 4. *Let $V = vv^T \in \mathcal{E}$. We define the following linear operator:*

$$\forall S \in \mathcal{S}_p, \; \mathcal{L}_v(S) = \mathrm{Diag}(SV) - S.$$

Then:

1. *Its kernel is the set of diagonal matrices.*

2. $\mathcal{L}_v^2 = -\mathcal{L}_v$.

3. *Its image is* $\{S \in \mathcal{S}_p : \operatorname{Diag} SV = 0\}$ *consequently* $\forall S \in \mathcal{S}_p,\ v^T \mathcal{L}_v(S) v = 0$.

4. $\mathcal{L}_v^{-1}(\mathcal{S}_p^+) = \mathcal{C}_V^*$.

Proof. 1. If D is a diagonal matrix, $DV = D$, so $\mathcal{L}_v(D) = 0$. If $\mathcal{L}_v(S) = 0$ then as $\operatorname{Diag}(SV)$ is diagonal, the non diagonal terms of $-S$ are null so S is diagonal.

2. Let $S \in \mathcal{S}_p$, as seen a line above, $\mathcal{L}_v(\operatorname{Diag}(SV)) = 0$, so by linearity of $\mathcal{L}_v$, $\mathcal{L}_v(\mathcal{L}_v(S)) = \mathcal{L}_v(-S)$.

3. Its image is the image of $-\mathcal{L}_v$, which is a linear projector, so it is the set $\{S \in \mathcal{S}_p : -\mathcal{L}_v(S) = S\}$, thus the result. Consequently, if S is in this set, $v^T S v = \operatorname{tr}(SV) = 0$, and $\mathcal{L}_v(S)$ is symmetric.

4. $\boxed{\subset}$ Let $S \in \mathcal{L}_v^{-1}(\mathcal{S}_p^+) = \{S \in \mathcal{S}_p : \mathcal{L}_v(S) \succcurlyeq 0\},\ U \in \mathcal{E}^*$, then as $U \succcurlyeq 0$, $\operatorname{tr}(\mathcal{L}_v(S)U) \geqslant 0$, and we remark $\operatorname{tr}(\operatorname{Diag}(SV)U) = \operatorname{tr}(\operatorname{Diag}(SV)\operatorname{Diag}(U)) = \operatorname{tr}(SV)$, so
$$0 \leqslant \operatorname{tr}(\mathcal{L}_v(S)U) = \operatorname{tr}(SV) - \operatorname{tr}(SU).$$
Hence $\{S \in \mathcal{S}_p : \mathcal{L}_v(S) \succcurlyeq 0\} \subset \mathcal{C}_V^*$.
$\boxed{\supset}$ We proceed by contradiction. Let $S \in \mathcal{S}_p$ so that $\exists x \in^n$ with $x^T \mathcal{L}_v(S) x < 0$.
$$\begin{aligned} x^T \mathcal{L}_v(S) x &= x^T \operatorname{Diag}(SV) x - x^T S x, \\ &= \sum_{i,j} x_i^2 S_{ij} - \sum_{i,j} x_i x_j S_{ij}. \end{aligned}$$
As $S_{ij} = S_{ji}$, $\sum_{i,j} x_i^2 S_{ij} = \frac{1}{2}\sum_{i,j}(x_i^2 + x_j^2) S_{ij}$, so
$$x^T \mathcal{L}_v(S) x = \frac{1}{2}\sum_{i,j}(x_i^2 + x_j^2 - 2x_i x_j) S_{ij} = \frac{1}{2}\operatorname{tr}(SQ),$$
with the matrix $Q = \left[(x_i - x_j)^2\right]_{i,j}$. So we have $\operatorname{tr}(SQ) < 0$. Let $T = V - Q = \left[1 - (x_i - x_j)^2\right]_{i,j}$, then $\operatorname{tr}(ST) > \operatorname{tr}(SV)$. Moreover $T \succcurlyeq 0$ because T is a diagonally dominant symmetric matrix with real non-negative diagonal entries [8], and $\operatorname{Diag}(T) = I_p$, thus $T \in \mathcal{E}^*$. We deduce that $S \in \mathcal{C}_V^*$, and finally by contradiction $\mathcal{C}_V^* \subset \mathcal{L}_v^{-1}(\mathcal{S}_p^+)$.

□

Remark. The fourth result in this Proposition justifies the optimality of our proof strategy.

We get

$$\mathbb{P}\left(\hat{V}^* \neq V\right) = \mathbb{P}\left(\mathcal{L}_v(\hat{\Sigma}) \not\succcurlyeq 0\right).$$

Since $\hat{\Sigma}$ is the empirical covariance, it is close to the covariance matrix Σ_v. We note $\|\cdot\|_{op}$ the operator norm defined by

$$\forall S \in \mathcal{S}_p, \; \|S\|_{op} = \sup_{\|x\|=1} \left|x^T S x\right|.$$

It is also the maximal absolute value of the eigenvalues of S. We have the following property:

Proposition 5. *If* $\left\|\mathcal{L}_v(\hat{\Sigma} - \Sigma_v)\right\|_{op} \leqslant \lambda$, *then* $\mathcal{L}_v(\hat{\Sigma}) \succcurlyeq 0$.

Proof. We remark that

$$\mathcal{L}_v(V) = \mathrm{Diag}(V^2) - V = pI_p - V$$

so

$$\mathcal{L}_v(\Sigma_v) = \frac{\lambda}{p}\mathcal{L}_v(V) = \lambda\pi_v,$$

with $\pi_v = \left(I_p - \frac{1}{p}vv^T\right)$ the orthogonal projector onto $(v)^\perp$. With the third item of the proposition 4, we can show that for all $x \in^p$ with norm $\|y\| = 1$,

$$x^T\mathcal{L}_v(\hat{\Sigma})x = (\pi_v x)^T\mathcal{L}_v(\hat{\Sigma})\pi_v x,$$

then

$$x^T\mathcal{L}_v(\hat{\Sigma})x = (\pi_v x)^T\mathcal{L}_v(\hat{\Sigma}-\Sigma_v)\pi_v x+\lambda\|\pi_v x\|^2 \geqslant -\|\pi_v x\|^2\left\|\mathcal{L}_v(\hat{\Sigma}-\Sigma_v)\right\|_{op}+\lambda\|\pi_v x\|^2,$$

then we use the hypothesis $\left\|\mathcal{L}_v(\hat{\Sigma} - \Sigma_v)\right\|_{op} \leqslant \lambda$:

$$\forall x \in^p, \; x^T\mathcal{L}_v(\hat{\Sigma})x \geqslant 0.$$

□

The result is:

$$\boxed{\mathbb{P}\left(\hat{V}^* \neq V\right) \leqslant \mathbb{P}\left(\left\|\mathcal{L}_v(\hat{\Sigma} - \Sigma_v)\right\|_{op} > \lambda\right).} \tag{1}$$

3.2 Decomposition

We now decompose the operator norm with triangular inequality:

$$\left\|\mathcal{L}_v(\hat{\Sigma}-\Sigma_v)\right\|_{op} \leqslant \left\|\mathrm{Diag}\left((\hat{\Sigma}-\Sigma_v)V\right)\right\|_{op} + \left\|\hat{\Sigma}-\Sigma_v\right\|_{op}.$$

We can further decompose the first term in this technical lemma:

Lemma 1. *There exists i.i.d. random variables $Y_1,\dots,Y_n$ following the standard multivariate normal distribution $\mathcal{N}_p(0,I_p)$, so that with $\hat{S}=\frac{1}{n}\sum_{k=1}^n Y_k Y_k^T$ the sample covariance,*

$$\left\|\mathrm{Diag}\left((\hat{\Sigma}-\Sigma_v)V\right)\right\|_{op} \leqslant \left(1+\lambda-\sqrt{1+\lambda}\right)\left|\frac{v^T(\hat{S}-I_p)v}{p}\right| + \sqrt{1+\lambda}\max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{S}-I_p)v\right|$$

Proof. The operator norm of a diagonal matrix is the maximum value of its coefficients's absolute value:

$$\left\|\mathrm{Diag}\left((\hat{\Sigma}-\Sigma_v)V\right)\right\|_{op} = \max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{\Sigma}-\Sigma_v)vv^Te_i)\right|,$$

and since $e_i^Tv=v_i=\pm 1$,

$$\left\|\mathrm{Diag}\left((\hat{\Sigma}-\Sigma_v)V\right)\right\|_{op} = \max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{\Sigma}-\Sigma_v)v)\right|.$$

We can remark that Σ_v is a positive definite matrix, thus there exists a square root $\Sigma_v^{1/2}$ that is also positive definite. So there also exists an inverse of the square root $\Sigma_v^{-1/2}$. Let $Y_1,\dots,Y_p$ be the random vectors defined by $Y_k=\Sigma_v^{-1/2}X_k$. Thus the new sample covariance matrix $\hat{S}=\frac{1}{n}\sum_{k=1}^p Y_kY_k^T$ verifies $\hat{\Sigma}=\Sigma_v^{1/2}\hat{S}\Sigma_v^{1/2}$. Then

$$\left\|\mathrm{Diag}\left((\hat{\Sigma}-\Sigma_v)V\right)\right\|_{op} = \max_{1\leqslant i\leqslant p}\left|e_i^T\Sigma_v^{1/2}(\hat{S}-I_p)\Sigma_v^{1/2}v)\right|.$$

With π_v defined as in the previous lemma, we have

$$\Sigma_v = \pi_v + (1+\lambda)\frac{vv^T}{p},$$

yet $\frac{vv^T}{p}$ is the orthogonal projector onto v, so

$$\Sigma_v^{1/2} = \pi_v + \sqrt{1+\lambda}\frac{vv^T}{p} = I_p + \left(\sqrt{1+\lambda}-1\right)\frac{vv^T}{p}.$$

We know that $\frac{vv^T}{p}e_i=\frac{v}{p}$, so

$$\forall 1\leqslant i\leqslant p,\ \Sigma_v^{1/2}e_i = e_i + \frac{\left(\sqrt{1+\lambda}-1\right)}{p}v \ \text{ and } \Sigma_v^{1/2}v=\sqrt{1+\lambda}v.$$

We compute:

$$\left\|\mathrm{Diag}\left((\hat{\Sigma}-\Sigma_v)V\right)\right\|_{op} = \max_{1\leqslant i\leqslant p}\left|\left(e_i + \frac{(\sqrt{1+\lambda}-1)}{p}v\right)^T(\hat{S}-I_p)(\sqrt{1+\lambda}v))\right|,$$

$$= \max_{1\leqslant i\leqslant p}\left|\sqrt{1+\lambda}e_i^T(\hat{S}-I_p)v + \left(1+\lambda-\sqrt{1+\lambda}\right)\frac{v^T(\hat{S}-I_p)v}{p}\right|.$$

We get the result with the triangular inequality. □

We can also decompose the second term:

Lemma 2. *There exists i.i.d. random variables $Y_1',\dots,Y_n'$ following the standard multivariate normal distribution $\mathcal{N}_{p-1}(0,I_{p-1})$, with $\hat{S}'=\frac{1}{n}\sum_{k=1}^n Y_k'Y_k'^T$ the sample covariance, and i.i.d. variables $Z_1',\dots,Z_n'$ following the distribution $\mathcal{N}(0,1)$, so that*

$$\left\|\hat{\Sigma}-\Sigma_v\right\|_{op} \leqslant \left\|\hat{S}'-I_{p-1}\right\|_{op} + \frac{1+\lambda}{n}\left|\sum_{k=1}^n (Z_k'^2-1)\right|.$$

Proof. We write the definition

$$\left\|\hat{\Sigma}-\Sigma_v\right\|_{op} = \max_{\|y\|=1}\left|x^T(\hat{\Sigma}-\Sigma_v)x\right|.$$

We need to separate the direction $v'=v/\sqrt{p}$ in the sphere,

$$\left\|\hat{\Sigma}-\Sigma_v\right\|_{op} = \max{}_{\theta\in[0,2\pi],\ x'\perp v',\|x'\|=1}\left|(\cos(\theta)x'+\sin(\theta)v')^T(\hat{\Sigma}-\Sigma_v)(\cos(\theta)x'+\sin(\theta)v')\right|.$$

We develop $\hat{\Sigma}=\frac{1}{n}\sum_{k=1}^n X_kX_k^T$, and we decompose $X_k = X_k' + \sqrt{1+\lambda}Z_k'v'$, with $Z_k' = \frac{X_k^Tv'}{\sqrt{1+\lambda}}$. Then $X_k' = \pi_v X_k \perp v$, Z_k' follows a standard normal distribution $\mathcal{N}(0,1)$. The formula becomes

$$\left\|\hat{\Sigma}-\Sigma_v\right\|_{op} = \frac{1}{n}\max_{\theta\in[0,2\pi],\ x'\perp v',\|x'\|=1}\left|\left(\sum_{k=1}^n((X_k'^Tx')^2-1)\right)\cos^2(\theta) + (1+\lambda)\left(\sum_{k=1}^n(Z_k'^2-1)\right)\sin^2(\theta)\right|.$$

We know that $\cos^2\theta+\sin^2\theta=1$, so the maximum over θ is the maximum of the absolute value of one of the two terms:

$$\left\|\hat{\Sigma}-\Sigma_v\right\|_{op} = \frac{1}{n}\max\left\{(1+\lambda)\left|\sum_{k=1}^n(Z_k'^2-1)\right|;\ \max_{x'\perp v',\|x'\|=1}\left|\sum_{k=1}^n(X_k'^Tx')^2-1\right|\right\}.$$

So

$$\left\|\hat{\Sigma}-\Sigma_v\right\|_{op} \leqslant \frac{1}{n}(1+\lambda)\left|\sum_{k=1}^n(Z_k'^2-1)\right| + \frac{1}{n}\max_{x'\perp v',\|x'\|=1}\left|\sum_{k=1}^n(X_k'^Tx')^2-1\right|.$$

Let $\phi : \{v\}^{\perp} \to^{p-1}$ be an isometry, then $Y_k' = \phi(X_k') \sim \mathcal{N}_{p-1}(0, I_{p-1})$ and the set

$$\{\phi(x') : x' \perp v',\ \|x'\| = 1\}$$

is just the $p-1$ dimensional sphere S^{p-1}. Moreover by the isometry property,

$$Y_k'^T(\phi(x')) = (\phi(X_k'))^T(\phi(x')) = X_k'^T x',$$

so we get with $\hat{S}' = \frac{1}{n}\sum_{k=1}^n Y_k' Y_k'^T$:

$$\max_{x' \perp v', \|x'\|=1} \left| \frac{1}{n} \sum_{k=1}^n (X_k'^T x')^2 - 1 \right| = \left\| \hat{S}' - I_{p-1} \right\|_{op},$$

and thus the result. □

We finally have the following decomposition:

$$\boxed{\begin{aligned} \left\| \mathcal{L}_v(\hat{\Sigma} - \Sigma_v) \right\|_{op} \leqslant & \sqrt{1+\lambda} \max_{1 \leqslant i \leqslant p} \left| e_i^T (\hat{S} - I_p) v \right| + \left(1 + \lambda - \sqrt{1+\lambda}\right) \left| \frac{v^T(\hat{S} - I_p)v}{p} \right| \\ & + \frac{1+\lambda}{n} \left| \sum_{k=1}^n (Z_k'^2 - 1) \right| + \left\| \hat{S}' - I_{p-1} \right\|_{op}. \end{aligned}} \tag{2}$$

3.3 Link with χ_n^2 tail

The Wishart distribution is often seen as a generalization of the χ_n^2 distribution. Let Z be a real random variable following the χ_n^2 distribution, we note

$$f_n(y) := \mathbb{P}\left(\frac{|Z - n|}{n} \geqslant y \right)$$

the tail of the χ_n^2 distribution. We can remark that for two real random variables $Z,\ Z'$, for all $\gamma \in (0,1)$, with $\gamma' = 1 - \gamma$,

$$\mathbb{P}\left(Z + Z' \geqslant y\right) \leqslant \mathbb{P}\left(Z \geqslant \gamma y\right) + \mathbb{P}\left(Z' \geqslant \gamma' y\right). \tag{3}$$

Concentration bounds for the first three terms of the decomposition (2) follow.

Lemma 3. *For all $y > 0$ we have*

1.

$$\mathbb{P}\left(\sqrt{1+\lambda} \max_{1 \leqslant i \leqslant p} \left| e_i^T (\hat{S} - I_p) v \right| \geqslant y \right) \leqslant 2p f_n \left(\frac{y}{\sqrt{(1+\lambda)p}} \right),$$

2.

$$\mathbb{P}\left(\left(1+\lambda-\sqrt{1+\lambda}\right)\left|\frac{v^T(\hat{S}-I_p)v}{p}\right|+\frac{1+\lambda}{n}\left|\sum_{k=1}^{n}(Z_k'^2-1)\right|\geqslant y\right)\leqslant 2f_n\left(\frac{y}{2(1+\lambda)-\sqrt{1+\lambda}}\right).$$

Proof. 1. We write the union bound

$$\mathbb{P}\left(\max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{S}-I_p)v\right|\geqslant\frac{y}{\sqrt{1+\lambda}}\right)\leqslant\sum_{1\leqslant i\leqslant p}\mathbb{P}\left(\left|e_i^T(\hat{S}-I_p)v\right|\geqslant\frac{y}{\sqrt{1+\lambda}}\right),$$
$$\leqslant\sum_{1\leqslant i\leqslant p}\mathbb{P}\left(\left|e_i^T(\hat{S}-I_p)v'\right|\geqslant\frac{y}{\sqrt{(1+\lambda)p}}\right),$$

with $v'=\frac{v}{\sqrt{p}}$. To express $e_i^T(\hat{S}-I_p)v'$ with χ_n^2 distribution, we use a polarization identity :

$$e_i^T(\hat{S}-I_p)v'=\frac{1}{4}\left((e_i+v')(\hat{S}-I_p)(e_i+v')-(e_i-v')(\hat{S}-I_p)(e_i-v')\right).$$

$$\left|e_i^T(\hat{S}-I_p)v'\right|\leqslant\frac{1}{4}\left(\left|(e_i+v')(\hat{S}-I_p)(e_i+v')\right|+\left|(e_i-v')(\hat{S}-I_p)(e_i-v')\right|\right).$$

We remark that $\frac{(e_i+v')(\hat{S}-I_p)(e_i+v')}{\|e_i+v\|^2}$ and $\frac{(e_i-v')(\hat{S}-I_p)(e_i-v')}{\|e_i-v\|^2}$ are two centered χ_n^2 variables. Then, we have two cases,

- if $v_i=1$ then $\|e_i+v'\|^2=2+\frac{2}{\sqrt{p}}$ and $\|e_i-v'\|^2=2-\frac{2}{\sqrt{p}}$,
- if $v_i=-1$ then $\|e_i+v'\|^2=2-\frac{2}{\sqrt{p}}$ and $\|e_i-v'\|^2=2+\frac{2}{\sqrt{p}}$,

in both cases, there exists Z and Z' two χ_n^2 variables so that

$$\left|e_i^T(\hat{S}-I_p)v'\right|\leqslant\frac{1}{2}\left(1+\frac{1}{\sqrt{d}}\right)\frac{|Z-n|}{n}+\frac{1}{2}\left(1-\frac{1}{\sqrt{d}}\right)\frac{|Z'-n|}{n}.$$

With $\eta=\frac{1}{2}\left(1+\frac{1}{\sqrt{d}}\right)\in(0,1)$, $\eta'=1-\eta=\frac{1}{2}\left(1-\frac{1}{\sqrt{d}}\right)$, we have

$$\left|e_i^T(\hat{S}-I_p)v'\right|\leqslant\eta\frac{|Z-n|}{n}+\eta'\frac{|Z'-n|}{n},$$

we can now use the fact (3),

$$\mathbb{P}\left(\left|e_i^T(\hat{S}-I_p)v'\right|\geqslant\frac{y}{\sqrt{(1+\lambda)p}}\right)\leqslant\mathbb{P}\left(\frac{|Z-n|}{n}\geqslant\frac{y}{\sqrt{(1+\lambda)p}}\right)+\mathbb{P}\left(\frac{|Z'-n|}{n}\geqslant\frac{y}{\sqrt{(1+\lambda)p}}\right),$$
$$\leqslant 2f_n\left(\frac{y}{\sqrt{(1+\lambda)p}}\right).$$

We then have the result with our previous union bound.

2. Let $\gamma \in (0,1)$, $\gamma' = 1-\gamma$, then by (3),

$$\mathbb{P}\left(\left(1+\lambda-\sqrt{1+\lambda}\right)\left|\frac{v^T(\hat{S}-I_p)v}{p}\right| + \frac{1+\lambda}{n}\left|\sum_{k=1}^{n}(Z_k'^2-1)\right| \geqslant y\right)$$

$$\leqslant \mathbb{P}\left(\left|\frac{v^T(\hat{S}-I_p)v}{p}\right| \geqslant \frac{\gamma y}{1+\lambda-\sqrt{1+\lambda}}\right) + \mathbb{P}\left(\frac{\left|\sum_{k=1}^{n}(Z_k'^2-1)\right|}{n} \geqslant \frac{\gamma' y}{1+\lambda}\right).$$

The vector v has norm $\sqrt{p}$, so $\frac{v^T\hat{S}v}{p} = \sum_{k=1}^{n}(v^TY_k)^2/p$ follows a χ_n^2 distribution, so

$$\mathbb{P}\left(\left|\frac{v^T(\hat{S}-I_p)v}{p}\right| \geqslant \frac{\gamma y}{1+\lambda-\sqrt{1+\lambda}}\right) = f_n\left(\frac{\gamma y}{1+\lambda-\sqrt{1+\lambda}}\right).$$

Idem for the second term,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{k=1}^{n}(Z_k'^2-1)\right| \geqslant \frac{\gamma' y}{1+\lambda}\right) = f_n\left(\frac{\gamma' y}{1+\lambda}\right).$$

We then take $\gamma' = \frac{1+\lambda}{2(1+\lambda)-\sqrt{1+\lambda}}$ to equalize the terms in f_n, and we get the result.

□

We focus on the last term of (2). We remind of the ϵ-net argument [12].

Proposition 6 (ϵ-net argument)**.** *Let $\mathcal{N}_\epsilon$ be a maximal ϵ-net of the d dimensional unit sphere S^d, i.e. a set of points in S^d that are separated from each other by a distance of at least ϵ and which is maximal with respect to set inclusion. Then we have the following results:*

1. *The family of balls of radius ϵ centered with points in $\mathcal{N}_\epsilon$ is a covering of the sphere.*
2. *For all $d \times d$ dimensional matrix,*

$$\|M\|_{op} \leqslant (1-2\epsilon)^{-1}\max_{x\in\mathcal{N}_\epsilon}\left|x^TMx\right|,$$

3. *$\mathcal{N}_\epsilon$ has cardinality at most $(1+\frac{2}{\epsilon})^d$.*

Proof. 1. Let $x \in S^d$, then if $d(x,\mathcal{N}_\epsilon) \geqslant \epsilon$ we can add x to $\mathcal{N}_\epsilon$ and still get an ϵ-net. It contradicts with the maximality of $\mathcal{N}_\epsilon$, so

$$x \in \bigcup_{x'\in\mathcal{N}_\epsilon} B(x',\epsilon).$$

2. Let $x_m = \arg\max_{x \in S^d} \left| x^T M x \right|$, let $x \in \mathcal{N}_\epsilon$ be the closest point to x_m in $\mathcal{N}_\epsilon$, then

$$(x_m + x)^T M (x_m - x) \leqslant \|x_m + x\| \, \|x_m - x\| \, \|M\|_{op}.$$

By the first property, $\|x_m - x\| \leqslant \epsilon$, and by triangular inequality $\|x_m + x\| \leqslant 2$, so

$$(x_m + x)^T M (x_m - x) \leqslant 2\epsilon \, \|M\|_{op}.$$

Moreover

$$(x_m + x)^T M (x_m - x) = x_m^T M x_m - x^T M x = \|M\|_{op} - x^T M x \geqslant \|M\|_{op} - \max_{x \in \mathcal{N}_\epsilon} \left| x^T M x \right|,$$

combining the two inequalities we get the result.

3. The balls of radius $\frac{\epsilon}{2}$ and center in $\mathcal{N}_\epsilon$ are disjointed, and included in a ball of radius $1 + \frac{\epsilon}{2}$ so the sum of their volumes is smaller than the volume of the big ball.

$$|\mathcal{N}_\epsilon| \left(\frac{\epsilon}{2}\right)^d \leqslant \left(1 + \frac{\epsilon}{2}\right)^d$$

so we have the result.

□

The tail bound for standard Wishart matrices follows:

Lemma 4. *For all $y > 0$, for all $\epsilon \in (0, 1/2)$,*

$$\mathbb{P}\left(\left\| \hat{S}' - I_{p-1} \right\|_{op} \geqslant y \right) \leqslant \left(1 + \frac{2}{\epsilon}\right)^{p-1} f_n((1 - 2\epsilon)y).$$

Proof. Let $\mathcal{N}_\epsilon$ be an ϵ-net of S^{p-1}, we use proposition 6,

$$\begin{aligned}\mathbb{P}\left(\left\| \hat{S}' - I_{p-1} \right\|_{op} \geqslant y \right) &\leqslant \mathbb{P}\left(\max_{x \in \mathcal{N}_\epsilon} \left| x^T (\hat{S}' - I_{p-1}) x \right| \geqslant (1 - 2\epsilon)y \right), \\ &\leqslant \sum_{x \in \mathcal{N}_\epsilon} \mathbb{P}\left(\left| x^T (\hat{S}' - I_{p-1}) x \right| \geqslant (1 - 2\epsilon)y \right).\end{aligned}$$

We remark that for all $x \in S^{p-1}$, $x^T (\hat{S}' - I_{p-1}) x = \frac{1}{n} \sum_{k=1}^n (Y_k'^T x)^2 - 1$ and we know $Y_k'^T x \sim \mathcal{N}(0, 1)$, so

$$\mathbb{P}\left(\left| x^T (\hat{S}' - I_{p-1}) x \right| \geqslant (1 - 2\epsilon)y \right) = f_n((1 - 2\epsilon)y),$$

then

$$\mathbb{P}\left(\left\| \hat{S}' - I_{p-1} \right\|_{op} \geqslant y \right) \leqslant |\mathcal{N}_\epsilon| \, f_n((1 - 2\epsilon)y),$$

and the proposition 6 gives us this inequality:

$$\mathbb{P}\left(\left\| \hat{S}' - I_{p-1} \right\|_{op} \geqslant y \right) \leqslant \left(1 + \frac{2}{\epsilon}\right)^{p-1} f_n((1 - 2\epsilon)y).$$

□

3.4 Concentration

Here is a sharp concentration bound for χ_n^2 variables (inspired by [3]):

Proposition 7 (Concentration bound for χ_n^2)**.** *Let Z be a real random variable following the χ_n^2 distribution, then for all $y > 0$,*

$$f_n(y) = \mathbb{P}\left(\frac{|Z-n|}{n} \geqslant y\right) \leqslant 2e^{-\frac{ny^2}{8(y\vee 1)}},$$

with the notation $a \vee b = \max\{a, b\}$.

Proof. We use the Markov-Chernoff bound:

$$\mathbb{P}\left(\frac{Z-n}{n} \geqslant y\right) \leqslant \inf_{t\in} e^{-n(y+1)t}\mathbb{E}\left[e^{tZ}\right],$$

moreover for $t \leqslant 1/2$, $\mathbb{E}\left[e^{tZ}\right] = (1-2t)^{-n/2}$, so

$$\mathbb{P}\left(\frac{Z-n}{n} \geqslant y\right) \leqslant \inf_{t\leqslant 1/2} e^{-\frac{n}{2}(2(y+1)t+\log(1-2t))},$$

and we derive the term in the exponential to find the minimum at $t_{\min} = \frac{1}{2}\left(1-\frac{1}{1+y}\right)$, and we finally compute

$$\mathbb{P}\left(\frac{Z-n}{n} \geqslant y\right) \leqslant e^{-\frac{n}{2}(y-\log(1+y))}.$$

The tail bound for $\mathbb{P}\left(\frac{Z-n}{n} \leqslant -y\right)$ is proved identically. We note that

$$f_n(y) \leqslant \mathbb{P}\left(\frac{Z-n}{n} \geqslant y\right) + \mathbb{P}\left(\frac{Z-n}{n} \leqslant -y\right),$$

so

$$f_n(y) \leqslant 2e^{\frac{n}{2}(\log(1+y)-y)}.$$

$$f_n(y) \leqslant 2e^{-n\frac{y}{4}\left(2-2\frac{\log(1+y)}{y}\right)},$$

We continue with the previous result, we have the following Taylor formula for the logarithm:

$$\log(1+y) = y - \frac{y^2}{2} + \frac{y^3}{3} - \int_0^y \frac{(y-t)^3}{(1+t)^4}dt.$$

In the last term the integrand is positive so

$$\log(1+y) - y \leqslant -\frac{y^2}{2} + \frac{y^3}{3},$$

thus

$$f_n(y) \leqslant 2e^{-n\frac{y^2}{4}\left(1-\frac{2y}{3}\right)}.$$

We can choose for any y this inequality or the previous, so

$$f_n(y) \leqslant 2e^{-n\frac{y\wedge y^2}{4}\left(\left(1-\frac{2y}{3}\right)\vee\left(2-2\frac{\log(1+y)}{y}\right)\right)},$$

moreover the functions $y \mapsto 1-\frac{2y}{3}$ and $y \mapsto 2-2\frac{\log(1+y)}{y}$ are respectively decreasing and increasing, and are equals at a point y_0, such that $1-2y_0/3 \geqslant 1/2$. So $\left(1-\frac{2y}{3}\right) \vee \left(2-2\frac{\log(1+y)}{y}\right) \geqslant 1/2$, and thus the result:

$$f_n(y) \leqslant 2e^{-\frac{ny}{8}(y\wedge 1)} = 2e^{-\frac{ny^2}{8(y\vee 1)}}.$$

□

Using this bound on the expressions of the lemmas 3 and 4, we get:

Lemma 5. *For all $y > 0$, and $p \geqslant 3$*

1. *The first expression of lemma 3 is bounded by the following:*

$$\mathbb{P}\left(\sqrt{1+\lambda}\max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{S}-I_p)v\right| \geqslant y\right) \leqslant 4e\,\exp\left(-\frac{ny^2}{8(1+\lambda)p\log p\left(y(p(1+\lambda))^{-1/2}\vee 1\right)}\right).$$

2. *The second expression of lemma 3 is bounded by the following:*

$$2f_n\left(\frac{y}{2(1+\lambda)-\sqrt{1+\lambda}}\right) \leqslant 4\exp\left(-\frac{ny^2}{16(1+\lambda)^2}\left(1-\frac{y}{3(1+\lambda)}\right)\right).$$

3. *With $c := \min_{\epsilon\in(0,1/2)}\frac{\log(1+2/\epsilon)}{(1-2\epsilon)^2} \approx 4.6$, we have the following bound for the expression in lemma 4:*

$$\mathbb{P}\left(\left\|\hat{S}'-I_{p-1}\right\|_{op} \geqslant y\right) \leqslant 2e\,\exp\left(-\frac{ny^2}{8cp(y\vee 1)}\right).$$

Proof. 1. By lemma 3, proposition 7, and factorizing by $\log p$,

$$\mathbb{P}\left(\sqrt{1+\lambda}\max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{S}-I_p)v\right| \geqslant y\right) \leqslant 4\exp\left(\log p\left(1-\frac{ny^2}{8(1+\lambda)p\log p\left(y((1+\lambda)p)^{-1/2}\vee 1\right)}\right)\right)$$

If the term in the exponential is positive, then the inequality is trivial, and if it is negative, we note that $\log p \geqslant \log 3 \geqslant 1$.

2. f_n is decreasing so $f_n\left(\frac{y}{2(1+\lambda)-\sqrt{1+\lambda}}\right) \leqslant f_n\left(\frac{y}{2(1+\lambda)}\right)$, then we use the second to last bound of the proof of proposition 7.

3. By lemma 4 and proposition 7, for all $\epsilon \in (0, 1/2)$, we have

$$\mathbb{P}\left(\left\|\hat{S}' - I_{p-1}\right\|_{op} \geqslant y\right) \leqslant 2\exp\left((p-1)\log(1+2/\epsilon)\left(1 - \frac{n(1-2\epsilon)^2 y^2}{8(p-1)\log(1+2/\epsilon)\left((1-2\epsilon)y \vee 1\right)}\right)\right)$$

We choose $\epsilon = \epsilon_m$ that maximizes $\frac{(1-2\epsilon)^2}{\log(1+2/\epsilon)}$. We note $c = \frac{\log(1+2/\epsilon_m)}{(1-2\epsilon_m)^2}$. Then we get

$$\mathbb{P}\left(\left\|\hat{S}' - I_{p-1}\right\|_{op} \geqslant y\right) \leqslant 2\exp\left((p-1)\log(1+2/\epsilon_m)\left(1 - \frac{ny^2}{8c(p-1)\,(y \vee 1)}\right)\right).$$

We conclude noting that $\log(1+2/\epsilon_m) \geqslant 1$ and $p - 1 \geqslant 1$.

□

3.5 Conclusion

We compile our previous results to get the final bound:

Theorem 1 (Upper bound)**.** *Let $\delta \in (0,1)$. There exists $c \approx 4.6$, so that if*

$$n \geqslant 8\left(\left(\frac{\lambda^2}{p(1+\lambda)}\right)^{1/4} \vee 1 + \sqrt{\frac{3(1+\lambda)}{p\log p}} + \sqrt{\frac{c}{\log p}}\right)^2 \frac{(1+\lambda)p\log p}{\lambda^2}\log\left((4+6e)\delta^{-1}\right),$$

then $\hat{V}^* = V$ with probability at least $1-\delta$. We can have a clearer bound if $p \geqslant 1+\lambda$. In this case, let $c_1 = \sqrt{3} + \sqrt{c} \approx 3.87$, and $c_2 = 6e + 4 \approx 20.3$, if

$$n \geqslant 8\left(1 + \frac{c_1}{\sqrt{\log p}}\right)^2 \frac{(1+\lambda)p\log p}{\lambda^2}\log\left(\frac{c_2}{\delta}\right),$$

then we have exact recovery with probability at least $1-\delta$.

Proof. Let α, β and γ be three positive real numbers so that $\alpha+\beta+\gamma = 1$, by the inequations (1), (2) and (3),

$$\begin{aligned}
\mathbb{P}\left(\hat{V}^* \neq V\right) &\leqslant \mathbb{P}\left(\left\|\mathcal{L}_v(\hat{\Sigma} - \Sigma_v)\right\|_{op} > \lambda\right), \\
&\leqslant \mathbb{P}\left(\sqrt{1+\lambda}\max_{1\leqslant i\leqslant p}\left|e_i^T(\hat{S} - I_p)v\right| \geqslant \alpha\lambda\right) \\
&\quad + \mathbb{P}\left(\left(1+\lambda-\sqrt{1+\lambda}\right)\left|\frac{v^T(\hat{S}-I_p)v}{p}\right| + \frac{1+\lambda}{n}\left|\sum_{k=1}^{n}(Z_k'^2 - 1)\right| \geqslant \beta\lambda\right) \\
&\quad + \mathbb{P}\left(\left\|\hat{S}' - I_{p-1}\right\|_{op} \geqslant \gamma\lambda\right).
\end{aligned}$$

Then, by lemma 5,

$$\mathbb{P}\left(\hat{V}^* \neq V\right) \leqslant 4e\ \exp\left(-\frac{n\alpha^2\lambda^2}{8(1+\lambda)p\log p\left(\alpha\lambda(p(1+\lambda))^{-1/2} \vee 1\right)}\right)$$
$$+ 4\exp\left(-\frac{n\beta^2\lambda^2}{16(1+\lambda)^2}\left(1-\frac{\beta\lambda}{3(1+\lambda)}\right)\right)$$
$$+ 2e\ \exp\left(-\frac{n\gamma^2\lambda^2}{8cp(\gamma\lambda \vee 1)}\right),$$

and using the facts $\alpha < 1$, $\frac{\lambda}{1+\lambda} < 1$, $\gamma < 1$ and $\lambda \vee 1 \leqslant 1 + \lambda$,

$$\mathbb{P}\left(\hat{V}^* \neq V\right) \leqslant 4e\ \exp\left(-\frac{n\lambda^2}{8(1+\lambda)p\log p}\left(\frac{\alpha^2}{\lambda(p(1+\lambda))^{-1/2} \vee 1}\right)\right)$$
$$+ 4\exp\left(-\frac{n\lambda^2}{8(1+\lambda)p\log p}\left(\frac{\beta^2\left(1-\beta/3\right)p\log p}{2(1+\lambda)}\right)\right)$$
$$+ 2e\ \exp\left(-\frac{n\lambda^2}{8(1+\lambda)p\log p}\left(\frac{\gamma^2\log p}{c}\right)\right).$$

We choose α, β and γ so that the terms in the exponentials are equals:

$$\frac{\alpha^2}{\lambda(p(1+\lambda))^{-1/2} \vee 1} = \frac{\beta^2\left(1-\beta/3\right)p\log p}{2(1+\lambda)} = \frac{\gamma^2\log p}{c} =: \frac{1}{F^2}$$

then, because $\alpha + \beta + \gamma = 1$, we have

$$F = \left(\frac{\lambda^2}{p(1+\lambda)}\right)^{1/4} \vee 1 + \sqrt{\frac{2(1+\lambda)}{(1-\beta/3)p\log p}} + \sqrt{\frac{c}{\log p}}.$$

So

$$\mathbb{P}\left(\hat{V}^* \neq V\right) \leqslant (6e+4)\exp\left(-\frac{n\lambda^2}{8(1+\lambda)F^2p\log p}\right).$$

If $(6e+4)\exp\left(-\frac{n\lambda^2}{8(1+\lambda)F^2p\log p}\right) \leqslant \delta$, we have exact recovery with probability at least $1-\delta$. Using $\beta < 1$, we get the first result. $\square$

Remark. The constant c_1 seems quite big, considering that with $p = 10^4$, $8\left(1+\frac{c'}{\sqrt{\log p}}\right)^2 \approx 41$, but this constant cannot be sharpened below $\sqrt{c} \approx 2.1$, where we have the same kind of problem. So our final bound seems as sharp as possible, with the method we used.

4 Lower Bound

A useful tool to obtain lower bounds on estimation problems is the KL divergence. Let P and Q be two probability distributions over the same measurable space, the Kullback-Leibler divergence between P and Q is defined by $KL(P||Q) = \mathbb{E}_{X\sim P}\left[\log\left(\frac{P(X)}{Q(X)}\right)\right]$ [9].

Lemma 6 (KL divergence of spiked Gaussian)**.** *Let $u, v \in \{-1,1\}^p$ be two vertices of the p-dimensional hypercube, we note $P_u \sim \mathcal{N}_p(0, \Sigma_u)$ and $P_v \sim \mathcal{N}_p(0, \Sigma_v)$ the associated multivariate spiked Gaussian distributions. Let $k = \sum_{i=1}^p \mathbf{1}_{u_i \neq v_i}$ be the Hamming distance between u and v. We have the following result:*

$$KL(P_u||P_v) = \frac{2\lambda^2}{1+\lambda}\left(k/p - (k/p)^2\right).$$

Proof. Let $X \sim \mathcal{N}_p(0, \Sigma_u)$, by definition,

$$\begin{aligned}
KL(P_u||P_v) =& \mathbb{E}\left[\log\left(\frac{P_u(X)}{P_v(X)}\right)\right] \\
=& \frac{1}{2}\mathbb{E}\left[X^T\Sigma_v^{-1}X - X^T\Sigma_u^{-1}X\right] \\
=& \frac{1}{2}\mathbb{E}\left[\operatorname{tr}\Sigma_v^{-1}XX^T - \operatorname{tr}\Sigma_u^{-1}XX^T\right] \\
=& \frac{1}{2}\operatorname{tr}\left(\left(\Sigma_v^{-1} - \Sigma_u^{-1}\right)\mathbb{E}\left[XX^T\right]\right) \\
=& \frac{\operatorname{tr}\left(\Sigma_v^{-1}\Sigma_u\right) - p}{2}.
\end{aligned}$$

We note that $\Sigma_v = \pi_v + (1+\lambda)\frac{vv^T}{p}$, as seen in lemma 1, hence

$$\Sigma_v^{-1} = \pi_v + \frac{1}{1+\lambda}\frac{vv^T}{p} = I_p + \left(\frac{1}{1+\lambda} - 1\right)\frac{vv^T}{p}$$

so

$$\Sigma_v^{-1} = I_p - \frac{\lambda}{1+\lambda}\frac{vv^T}{p}.$$

So we develop

$$\Sigma_v^{-1}\Sigma_u = I_p - \frac{\lambda}{1+\lambda}\frac{vv^T}{p} + \lambda\frac{uu^T}{p} - \frac{\lambda^2}{1+\lambda}\frac{vv^Tuu^T}{p^2}.$$

We remark that $v^Tu = p - 2k$, then $\operatorname{tr} vv^Tuu^T = (v^Tu)^2 = p^2(1-2k/p)^2$. We also know that $\operatorname{tr}(uu^T) = \operatorname{tr}(vv^T) = p$, so

$$\operatorname{tr}\left(\Sigma_v^{-1}\Sigma_u\right) = p - \frac{\lambda}{1+\lambda} + \lambda - \frac{\lambda^2}{1+\lambda}(1-2k/p)^2.$$

As $\lambda - \frac{\lambda}{1+\lambda} = \frac{\lambda^2}{1+\lambda}$ we get

$$\operatorname{tr}\left(\Sigma_v^{-1}\Sigma_u\right) - p = \frac{\lambda^2}{1+\lambda}\left(1 - (1-2k/p)^2\right) = \frac{4\lambda^2}{1+\lambda}\left(k/p - (k/p)^2\right)$$

and thus the result. $\square$

We then deduce a lower bound using Fano's lemma [6]:

Theorem 2 (Lower bound)**.** *Let $\delta \in (0,1)$. If*

$$n \leqslant \frac{(1-\delta) - \frac{\log 2}{\log p}}{1 - \frac{2}{p}} \frac{1+\lambda}{4\lambda^2} p \log p,$$

then any estimator of $v \in \{-1,1\}^p$ makes an error with probability at least δ.

Proof. Let V_k be a subset of the hypercube $\{-1,1\}^p$, with cardinality r, so that for all $v, v' \in V$, the Hamming distance between v and v' is smaller than the integer k. Let $\hat{v}$ be an estimation of v with the i.i.d sample vectors $X_1, \ldots, X_n \sim P_v$, then by Fano's inequality,

$$\max_{v \in V_k} P_v^{\otimes n}(\hat{v} \neq v) \geqslant 1 - \frac{\log 2 + \max_{u,v \in V_k} KL(P_u^{\otimes n} || P_v^{\otimes n})}{\log(r-1)}.$$

Using lemma 6, we compute

$$KL(P_u^{\otimes n} || P_v^{\otimes n}) = nKL(P_u || P_v) \leqslant \frac{2\lambda^2 n}{1+\lambda} \left(k/p - (k/p)^2\right).$$

We can say that for our general estimation problem, as the performance of the estimation is independent of the choice of v, then

$$\mathbb{P}(\hat{v} \neq v) \geqslant \max_{v \in V_k} P_v^{\otimes n}(\hat{v} \neq v) \geqslant 1 - \frac{\log 2}{\log(r-1)} - \frac{2\lambda^2 n}{(1+\lambda)p} \frac{k}{\log(r-1)} (1 - k/p).$$

To maximize r, we need to choose V_k as a Hamming ball in the hypercube. Moreover, the ball that minimizes $k/\log(r-1)$ is V_2. So we choose $V_2 = \{v^0, v^1, \ldots, v^p\}$ so that $v_i^j = 1$ if $i \neq j$ and $v_i^i = -1$. Then $r = p+1$. Finally,

$$\mathbb{P}(\hat{v} \neq v) \geqslant 1 - \frac{\log 2}{\log p} - \frac{4\lambda^2 n}{(1+\lambda) p \log p} (1 - 2/p).$$

□

Remark. We can get a slightly better bound replacing $\log 2$ by $h(P)$ with $P \geqslant \mathbb{P}(\hat{v} \neq v)$ obtained by our upper bound and $h(x) = -x \log x - (1-x)\log(1-x)$.

5 An Algorithm

We show an example of algorithm that work for this particular optimization problem. It is based on projected gradient descent. Other algorithms for SDP optimization can be found in the book from Boyd, Vanderbergue [4], or [13].

5.1 Some initial results

Stopping step. We have a guarantee to finish any algorithm with a simple stopping step, simply taking the signs of the matrix as entries for the matrix $\hat{V}$, when our approximation matrix is close enough to its sign matrix:

Proposition 8. *For any matrix $Y \in \mathcal{S}_p$,*

$$Y \succcurlyeq 0,\ Y \in \{\pm 1\}^{p\times p},\ \mathrm{Diag}(Y) = I_p \ \Leftrightarrow\ Y = yy^T,\ y \in \{\pm 1\}^p.$$

Proof. $\boxed{\Leftarrow}$ $yy^T \in \mathcal{S}_p^+ \cap \{\pm 1\}^{p\times p}$.

$\boxed{\Rightarrow}$ If the rank of Y is 1, then $Y = yy^T$ with $y^i n^p$, thus the diagonal entries of Y are of the form $Y_{ii} = y_i^2 = \pm 1 = 1$ so $y_i = \pm 1$. Let's suppose that $\operatorname{rank} Y > 1$. Then there exists i, j, k, l so that the submatrix of Y with rows $i < j$ and columns $k < l$ is of rank 2. Switching to the transpose matrix, we can assume that it is of the form

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},\quad \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}.$$

- If $i = k$ and $j = l$, the submatrix should be symmetric, which is contradictory.
- If $i = k$ and $j \neq l$ or the contrary, we can get the symmetric submatrices of rows and columns i, j, l or i, j, k that follow (they need to have a diagonal with ones):

  $$\begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix},\quad \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix},$$

  but neither are positive semidefinite.
- If $i \neq k$ and $j \neq l$ then the submatrix of rows and columns (i, j, k, l) will be of the form

  $$\begin{pmatrix} 1 & \cdot & 1 & -1 \\ \cdot & 1 & 1 & 1 \\ 1 & 1 & 1 & \cdot \\ -1 & 1 & \cdot & 1 \end{pmatrix},\quad \begin{pmatrix} 1 & \cdot & -1 & 1 \\ \cdot & 1 & -1 & -1 \\ -1 & -1 & 1 & \cdot \\ 1 & -1 & \cdot & 1 \end{pmatrix}$$

  and we can verify that it is then impossible to get such a matrix that is semidefinite positive. Thus the contradiction.

$\square$

Angles between $V - I_p$ and $\partial \mathcal{C}_V^*$.

Lemma 7. *We have the following bounds for angles between $V - I_p$ and $\partial \mathcal{C}_V^*$:*

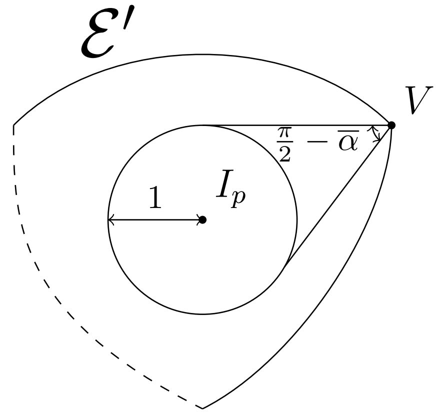
1. *The maximal angle between $V - I_p$ and $\partial \mathcal{C}_V^*$ is smaller than*

$$\overline{\alpha} = \frac{\pi}{2} - \arcsin\left(\frac{1}{\sqrt{p(p-1)}}\right).$$

2. *The minimal angle between $V - I_p$ and $\partial \mathcal{C}_V^*$ is greater than*

$$\underline{\alpha} = \arcsin\left(\frac{1}{\sqrt{p-1}}\right).$$

Proof. 1. We consider $\{\mathrm{Diag}\, M = I_p\}$.



The Frobenius distance between I_p and the boarder of the convex set $\mathcal{E}'$ is greater than its distance to the boarder of the positive semidefinite cone, which is 1. So the convex closure of $(B(I_p, 1) \cap \{\mathrm{Diag}\, M = I_p\}) \cup \{V\}$ is included in $\mathcal{E}'$. The angle between $V - I_p$ and the border of its normal cone at V is then smaller than

$$\overline{\alpha} = \arccos\left(\frac{1}{\|V - I_p\|}\right) = \frac{\pi}{2} - \arcsin\left(\frac{1}{\sqrt{p(p-1)}}\right).$$

Thus the maximal angle between $\partial \mathcal{C}_V^*$ and $V - I_p$ is smaller than this angle.

2. $S \notin \mathcal{C}_V^* \Rightarrow \mathcal{L}_v(S) \not\succeq 0$. But

$$\mathcal{L}_v\left(\frac{V - I_p}{p}\right) = \mathcal{L}_v\left(\frac{V}{p}\right) = I_p - \frac{V}{p} = \pi_v$$

is the orthogonal projection on $\{v\}^\perp$, and $\mathcal{L}_v(S)v = 0$ too. So $\left\|\mathcal{L}_v\left(\frac{V-I_p}{p}\right) - \mathcal{L}_v(S)\right\|_{op} > 1$. Thus

$$\left\|\frac{V - I_p}{p} - S\right\| \geqslant \frac{\left\|\mathcal{L}_v\left(\frac{V-I_p}{p}\right) - \mathcal{L}_v(S)\right\|}{|||\mathcal{L}_v|||} > \frac{1}{|||\mathcal{L}_v|||}.$$

Let S be the projection of $\frac{V-I_p}{p}$ on $\partial\mathcal{C}_V^*$, then $S \perp (V - I_p)/p - S$ since $\mathcal{C}_V^*$ is a cone centered at 0. Then the angle between S and $V - I_p$ is

$$\arcsin\left(\frac{\left\|\frac{V-I_p}{p} - S\right\|}{\left\|\frac{V-I_p}{p}\right\|}\right) > \arcsin\left(\frac{1}{\left\|\frac{V-I_p}{p}\right\| |||\mathcal{L}_v|||}\right) \geqslant \arcsin\left(\frac{1}{\left\|\frac{V-I_p}{p}\right\| |||\mathcal{L}_v|||}\right).$$

First, $\|V - I_p\| = \sqrt{p(p-1)}$. We know that diagonal matrices are in the kernel of $\mathcal{L}_v$ so

$$|||\mathcal{L}_v||| = \sup_{S \in \mathcal{S}_p} \frac{\|\mathcal{L}_v(S)\|}{\|S\|} = \sup_{\mathrm{Diag}(S)=0} \frac{\|\mathcal{L}_v(S)\|}{\|S\|}.$$

Let S be a matrix with a diagonal filled with zeros. We can separate $Lv(S)$ in its diagonal part and its anti-diagonal part:

$$\|\mathcal{L}_v(S)\|^2 = \|\mathrm{Diag}(SV)\|^2 + \|S\|^2.$$

By symmetry we choose $v = (1, \ldots, 1)$,

$$\begin{aligned}
\|\mathrm{Diag}(SV)\|^2 &= \mathrm{tr}\left((\mathrm{Diag}(SV)^2)\right) \\
&= \sum_{k=1}^{p}\left(\sum_{i \neq k} S_{ik}\right)^2 = \sum_{k=1}^{p}\sum_{i \neq k}\sum_{j \neq k} S_{ik}S_{jk} \\
&\leqslant \frac{1}{2}\sum_{k=1}^{p}\sum_{i \neq k}\sum_{j \neq k} S_{ik}^2 + S_{jk}^2 \\
&\leqslant (p-1)\sum_{k=1}^{p}\sum_{i \neq k} S_{ik}^2 = (p-1)\|S\|^2
\end{aligned}$$

So $|||\mathcal{L}_v||| = \sqrt{p}$. Finally the minimal angle between $V - I_p$ and $\partial\mathcal{C}_V^*$ is greater than

$$\underline{\alpha} = \arcsin\left(\frac{1}{\sqrt{p-1}}\right).$$

$\square$

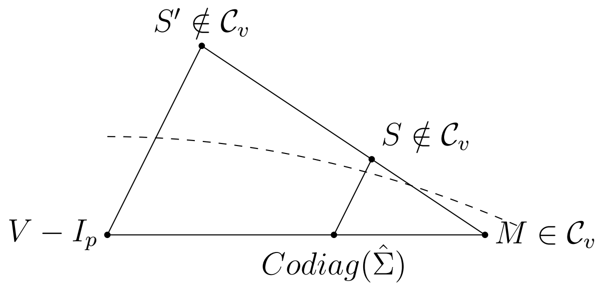
Angle of $\hat{\Sigma}$ with regards to $V - I_p$ and $\partial\mathcal{C}_V^*$. We proved that with high probability $\left\|\mathcal{L}_v\left(\hat{\Sigma} - \sigma_v\right)\right\|_{op} \leqslant \lambda$, which implies that $\hat{\Sigma}$ is in the cone $\mathcal{C}_V^*$. Then, also with high probability, $\left\|\mathcal{L}_v\left(\hat{\Sigma} - \sigma_v\right)\right\|_{op} \leqslant (1-\epsilon)\lambda$.

Lemma 8. *We suppose that* $\left\|\mathcal{L}_v\left(\hat{\Sigma} - \sigma_v\right)\right\|_{op} \leqslant (1-\epsilon)\lambda$, *then we have the following:*

1. *The angle between* $\mathrm{Codiag}(\hat{\Sigma}) = \hat{\Sigma} - \mathrm{Diag}(\hat{\Sigma})$ *and* $V - I$ *is smaller than* $(1-\epsilon)\overline{\alpha}$.
2. *The minimal angle between* $\mathrm{Codiag}(\hat{\Sigma})$ *and* $\partial\mathcal{C}_V^*$ *is greater than* $\epsilon\underline{\alpha}$

Proof. 1. $\hat{\Sigma}$ is in the image of the cone by an $(1-\epsilon)$ homotecy around the axis $V - I$.

2. Let $S \notin \mathcal{C}_V^*$. As $\mathrm{Codiag}\,\hat{\Sigma}$ is in the contracted cone there exists a matrix
$$M = \frac{1}{1-\epsilon}\left[\mathrm{Codiag}\,\hat{\Sigma} - (V - I_p)\right] + V - I_p$$



Then $M \in \mathcal{C}_V^*$ and we define $S' = M + \frac{1}{\epsilon}(S - M)$ (Thales figure), so that
$$S - \mathrm{Codiag}\,\hat{\Sigma} = \epsilon(S' - (V - I))$$
and $S' \notin \mathcal{C}_V^*$. If it was the case then by convexity S would be in $\mathcal{C}_V^*$ too. So the minimal angle between $\mathrm{Codiag}\,\hat{\Sigma}$ and $\partial\mathcal{C}_V^*$ is greater than $\epsilon\underline{\alpha}$.

□

5.2 Projected gradient

Algorithm. Let $\pi_{\mathcal{E}'}$ be a projection, with an error δ. Let γ be the gradient step size, and $\mathrm{sign}(Y)$ be the matrix with entries being ± 1 whether the sign of the corresponding entry of Y is positive or not.

The projected gradient algorithm is the following:

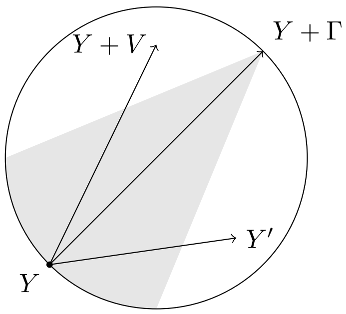
Algorithm 1 Calculate $\hat{V}^* \in \arg\max_{U \in \mathcal{E}^*} \operatorname{tr}(\hat{\Sigma} U)$

Require: $\hat{\Sigma} \succcurlyeq 0 \wedge \gamma > 0$
Ensure: $Y \in \arg\max_{U \in \mathcal{E}^*} \operatorname{tr}(\hat{\Sigma} U)$
  $\Gamma = \gamma \frac{\operatorname{Codiag} \hat{\Sigma}}{\|\operatorname{Codiag} \hat{\Sigma}\|}$
  $Y \leftarrow I_p + \Gamma$
  while $\operatorname{sign}(Y) \not\succcurlyeq 0$ **do**
    $Y \leftarrow \pi_{\mathcal{E}'}(Y + \Gamma)$
  end while
  $\hat{V}^* \leftarrow \operatorname{sign}(Y)$

Proposition 9. *The number of gradient steps is smaller than*

$$N_0(\delta) = \frac{\sqrt{p(p-1)}}{\gamma \sin(\epsilon \underline{\alpha}) \cos((1-\epsilon)\overline{\alpha} - \underline{\alpha}) - \delta}.$$

Proof. We first suppose the projection is an oracle projection with zero error. At one step, starting from the point Y, going to Y', the angle between $Y - Y'$ and $\Gamma + Y - Y'$ should be obtuse, because if not, there exists a point on the segment $[Y, Y'] \subset \mathcal{E}'$ that is closer to $Y + \Gamma$ than Y'. So Y' is in the ball of diameter $[Y, Y + \Gamma]$. Let $U = Y' - Y$. The projection of $Y + \Gamma$ on $\mathcal{E}'$ is in the direction $\Gamma - U$, so $\Gamma - U$ is in the normal cone of $\mathcal{E}'$ at the point Y', thus $Y' = V$ or $\Gamma - U \notin \mathcal{C}_V^*$. In the first case the algorithm finishes, and in the second case Y' cannot be in the set $\Gamma + Y - \mathcal{C}_V^*$, which is the grey area in the following figure.



So the maximal angle between U and Γ is smaller than $\frac{\pi}{2} - \epsilon\underline{\alpha}$. Then the angle between V and U is at most $(1-\epsilon)\overline{\alpha} + \frac{\pi}{2} - \epsilon\underline{\alpha}$. We can also see that $\|U\| \geqslant \sin(\epsilon\underline{\alpha})\gamma$.

Finally,

$$U^T\left(\frac{V-I_p}{\|V-I_p\|}\right) \geqslant \|U\| \sin\left((1-\epsilon)\overline{\alpha}+\frac{\pi}{2}-\epsilon\underline{\alpha}\right)$$
$$\geqslant \sin(\epsilon\underline{\alpha})\cos((1-\epsilon)\overline{\alpha}-\epsilon\underline{\alpha})\gamma.$$

So we have a guarantee that at each step, the oracle algorithm gets closer to the solution by this constant. Knowing that the starting point is I_p, if we add the error of approximation δ of the projection, we get the result. □

Dykstra's projection. As the convex set $\mathcal{E}'$ is the intersection of an affine set and $\mathcal{S}_p^+$, we can use Dykstra's projection algorithm for computing the projection on $\mathcal{E}'$.

The projection algorithm is the following:

Algorithm 2 Calculate $\pi_{\mathcal{E}'}(S)$

Require: $(S \in \mathcal{S}_p) \wedge (\delta > 0)$
Ensure: $d(S', \mathcal{E}') < \delta$
  $Y \leftarrow S$
  $Z \leftarrow \pi_{\mathcal{S}_p^+}(S)$
  while $\|\mathrm{Diag}(Z) - I_p\| \geqslant p\delta$ **do**
    $Y \leftarrow Y - \mathrm{Diag}\, Z + I_p$
    $Z \leftarrow \pi_{\mathcal{S}_p^+}(Y)$
  end while
  $S' = Z - \mathrm{Diag}(Z) + I_p$

After a gradient step, the distance to the projection on $\mathcal{E}'$ is smaller than γ, so the number of steps n of Dykstra's algorithm should guarantee that we get from a distance smaller than γ to a distance smaller than δ.

Proposition 10. *The number of steps of Dykstra algorithm needed to get from a distance to $\mathcal{E}'$ smaller than γ, to a distance smaller than δ, is smaller than*

$$n_0(\delta) = p \log\left(\frac{\gamma}{\delta}\right).$$

Proof. To check whether the projection method leads to a point close enough to $\mathcal{E}'$, we should first compute a lower bound for the angle between the two convex sets. Let $S \in \mathcal{E}'$, and let θ be the angle between S and its projection in $\{\mathrm{Diag}\, M = 0\}$. Then

$$\sin^2\theta = \frac{\mathrm{tr}(S\,\mathrm{Diag}\,S)}{\mathrm{tr}(S^2)} = \frac{\mathrm{tr}\left(\mathrm{Diag}(S)^2\right)}{\mathrm{tr}\,S^2} = \frac{p}{\mathrm{tr}(S^2)}.$$

Let $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ the eigenvalues of S counted with their multiplicity.

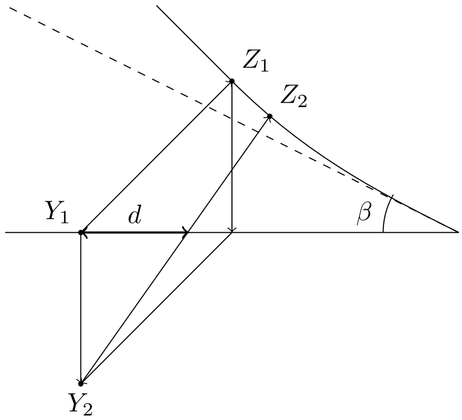
$$\operatorname{tr} S = \sum_i \lambda_i = p$$

So

$$\operatorname{tr} S^2 = \sum_i \lambda_i^2 \leqslant \left(\sum_i \lambda_i\right)^2 = p^2,$$

thus $\sin^2\theta \geqslant \frac{1}{p}$. So $\arcsin(1/\sqrt{p})$ is the minimal angle between $\mathcal{S}_p^+$ and $\{\operatorname{Diag} M = I_p\}$.

The algorithm can be shown with the following drawing:



Let L be the distance between Y_1 and the projection point, we can prove that $d \geqslant L\sin(\beta)^2$. The algorithm multiplies the distance to $\mathcal{E}'$ by at least $\cos(\beta)^2$ at each step. We also know that $\cos(\beta)^2 \geqslant 1 - 1/p$.

So the number n of steps of the algorithm is smaller than $-\frac{\log(\gamma/\delta)}{\log(1-1/p)}$, thus the result. $\square$

Complexity bound. We can conclude this section with the following theorem:

Theorem 3. *The algorithm has a complexity of*

$$p^5 \frac{\log(\gamma/delta)}{\gamma\epsilon\sin(\epsilon\pi/2)/\sqrt{p} - \delta}.$$

If we chose $\gamma \simeq p^{3/2}$ and $\delta = 1$, we get a single gradient step and a thus a complexity of $O(p^4\log(p))$.

Proof. Each gradient step has complexity p^2 because it is a sum of matrices, and a projection on $\mathcal{S}_p^+$ has complexity p^3. Thus the projected gradient algorithm has a complexity of $p^3 n_0(\delta) N_0(\delta)$. The result comes directly from the propositions 9 and 10. $\square$

6 Conclusion and remarks

We have proved that for our method, if

$$n \geqslant C_1(\delta)\frac{1+\lambda}{\lambda^2}p\log p,$$

then $\mathbb{P}\,(\text{exact recovery}) \leqslant \delta$. We have also showed that if

$$n < C_2(\delta)\frac{1+\lambda}{\lambda^2}p\log p,$$

then $\mathbb{P}\,(\text{exact recovery}) > \delta$. It expresses that our relaxation method is optimal, up to a constant, for finding a solution to this NP-hard problem with high probability in polynomial time. Each bound in this document is non-asymptotic, so that it matches a real use in computational statistics, though, the asymptotic versions of these results in λ, p or δ may be interesting for further analysis.

7 References

[1] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. February 2012.

[2] Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. Exact recovery in the ising blockmodel. *CoRR*, abs/1612.03880, 2016.

[3] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[5] Marcel K. de Carli Silva and Levent Tunçel. Vertices of Spectrahedra arising from the Elliptope, the Theta Body, and Their Relatives. September 2013.

[6] Robert M. Fano. *Transmission of Information: A Statistical Theory of Communication*. The MIT Press, 1961.

[7] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

[8] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. Johns Hopkins University Press, 1996.

[9] Solomon Kullback. *Information theory and statistics.* John Wiley and Sons, 1959.

[10] M. Y. Mo. The rank 1 real Wishart spiked model. January 2011.

[11] Jaehyun Park and Stephen Boyd. A semidefinite programming method for integer convex quadratic minimization. *Optimization Letters*, 12(3):499–518, 2018.

[12] Terence Tao. Topics in random matrix theory. 2011.

[13] Zaiwen Wen, Donald Goldfarb, and Wotao Yin. Alternating direction augmented lagrangian methods for semidefinite programming. *Mathematical Programming Computation*, 2(3):203–230, Dec 2010.

[14] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A(1-2):32–52, 1928.