ROBUST MEAN ESTIMATION ON THE MANIFOLD OF SYMMETRIC POSITIVE DEFINITE MATRICES

Alexandre Monti

Supervised by Yun Ho and Prof. Victor Panaretos

Contents

1 Introduction

The estimation of the mean is a fundamental problem in statistical analysis, providing crucial insights into the central tendencies of data distributions. Traditional mean estimation methods, such as the sample mean, assume that the data follow a Gaussian distribution and are devoid of outliers or influential observations. However, real-world data often deviate from these ideal assumptions, necessitating the development of robust estimation techniques.

In recent years, there has been growing interest in robust mean estimation[1,2,3]. This area of research addresses the challenge of estimating the mean of a distribution when the underlying data is not normally distributed or contaminated with outliers[4].

Robust mean estimation on symmetric positive definite matrices involves the development of statistical methodologies that are resilient to the presence of outliers or deviations from Gaussian assumptions. These methodologies aim to provide more accurate and reliable estimates of the mean, even when faced with challenging data scenarios.

The motivation behind robust mean estimation on SPD matrices stems from the limitations of classical estimators when applied to non-Gaussian or contaminated data. Outliers or influential observations can significantly impact the estimated mean, leading to biased results and incorrect inferences. By adopting robust approaches, researchers can obtain more robust estimates that are less sensitive to extreme values or departures from the Gaussian assumption[5].

The objective of this semester project is to investigate and develop robust mean estimation methods adapted specifically to SPD matrices. We will explore various robust estimation techniques and we will make simulations with one particular technique which is the most computation-friendly compared to the others.

By exploring the field of robust mean estimation on symmetric positive definite matrices, this project contributes to the advancement of statistical methodologies and provides valuable tools for researchers and practitioners working with complex data structures. The robust technique developed in this project has the potential to enhance data analysis in various domains where the mean of a symmetric positive definite matrix is of interest.

In the subsequent chapters, we will review the relevant literature[6], present the theoretical foundations of robust mean estimation on symmetric positive definite matrices, propose robust estimators, and discuss the implications and potential applications of our findings.

[1] Exponential Concentration for Geometric-Median-of-Means in Non-Positive Curvature Spaces, Ho Yun, Byeong U. Park, arXiv:2211.17155

[2] Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression, Olivier Catoni, Ilaria Giulini, arXiv:1712.02747

[3] Geometric means in a novel vector space structure on symmetric positive-definite matrices, Vincent Arsigny, Pierre Fillard, Xavier Pennec, Nicholas Ayache, http://www-sop.inria.fr/asclepios/Publications/Arsigny/arsigny_siam_tensors.pdf

[4] Mean estimation and regression under heavy-tailed distributions — a survey, Gabor Lugosi and Shahar Mendelson, arXiv:1906.04280

[5] Robust optimization and inference on manifolds, Lizhen Lin, Drew Lazar, Bayan Saparpabayeva and David Dunson, arXiv:2006.06843

[6] Riemannian geometric statistics in medical image analysis, Xavier Pennec, Stefan Sommer, Tom Fletcher, Elsevier, 2020.

2 Robust mean estimation on Euclidean spaces

We begin by exploring and defining the main concepts underlying robust mean estimation on Euclidean spaces. We introduce various techniques for robust mean estimation in one dimension and in multi-dimensions in order to be able to generalize this framework to manifolds later in this report.

2.1 One dimension case

We begin with the simplest case. Let $X_1, ..., X_n$ be an i.i.d. sample of real random variables with $\mathbb{E}[X_1] = \mu$. First, let's recall the definition of an estimator :

Definition : An estimator $\hat{\mu} = \hat{\mu}(X_1, ..., X_n)$ is a measurable function of $X_1, ..., X_n$.

One important feature of an estimator is the quality of the estimator. There exists multiple ways for measuring the quality of an estimator. The most common way to do this is to compute the mean-squared error : $\mathbb{E}[(\hat{\mu} - \mu)^2]$ but as for the empirical mean, the mean-squared error may be misleading. For our purpose, we prefer estimators $\hat{\mu}$ that are close to μ with high probability, i.e. $\mathbb{P}(|\hat{\mu} - \mu| > \epsilon) \leq \delta$ for a given sample size n and a given confidence parameter $\delta \in (0, 1)$. Thus, our purpose is to find the smallest $\epsilon = \epsilon(n, \delta)$ satisfying the last inequality. One important feature of this measurement of quality is that it is a non-asymptotic criterion, i.e. a quantitative criterion. Our aim is to find quantitative estimates for the mean which are moreover robust. We have a criterion to ensure that a mean estimator is quantitative :

Definition : A mean estimator $\hat{\mu}$ is *L-sub-Gaussian* if there exists a constant $L > 0$ such that for all sample sizes n we have

$$\mathbb{P}\left(|\hat{\mu} - \mu| \leq \frac{L\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right) \geq 1 - \delta$$

Our goal is to find estimators with sub-Gaussian error rate in the univariate as well as in the multivariate case. We present here two different estimators which have sub-Gaussian error rate for all distributions with finite variance.

2.1.1 The median-of-means estimator

Definition : Let $1 \leq k \leq n$ and partition $[n] = \{1, ..., n\}$ into k blocks $B_1, ..., B_k$, each of size $|B_i| \geq \lfloor n/k \rfloor \geq 2$. Given $X_1, ..., X_n$, compute the sample mean in each block $Z_j := \frac{1}{|B_j|}\sum_{i \in B_j} X_i$ and define the median-of-means estimator by $\hat{\mu}_n = M(Z_1, ..., Z_k)$ where $M(X_1, ..., X_n)$ is the median.

For this estimator, we have the following theorem :

Theorem : Let $X_1, ..., X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let m, k be positive integers and assume that $n = mk$. Then the median-of-means estimator $\hat{\mu}_n$ with k blocks satisfies

$$\mathbb{P}\left(|\hat{\mu}_n - \mu| > \sigma\sqrt{4/m}\right) \leq e^{-k/8}$$

In particular, for any $\delta \in (0, 1)$, if $k = \lceil 8\log(1/\delta) \rceil$, then we have

$$\mathbb{P}\left(|\hat{\mu}_n - \mu| \leq \sigma\sqrt{\frac{32\log(1/\delta)}{n}}\right) \geq 1-\delta$$

This theorem is particularly useful because it shows that the median-of-means estimator is L-sub-Gaussian for $L = 8$ and for all distribution with finite variance which is quite impressive ! However we can note that the sub-Gaussianity of $\hat{\mu}_n$ depends on the confidence level δ as the number of blocks k is defined depending on δ but we will see later that the sub-Gaussianity cannot be independent of the confidence level if we only assume that the variance is finite.

2.1.2 Catoni's estimator

The idea of Catoni to construct another mean estimator was to go from the formal definition of the empirical mean and to modify it. More precisely, the empirical mean $\hat{\mu}_n$ is the solution $y \in \mathbb{R}$ of the equation

$$\sum_{i=1}^{n}(X_i - y) = 0$$

The problem with this equation is that it gives the same "weight" to outliers than all other data. The idea of Catoni was then to modify the left-hand side of the equation in order to "down-weight" the outliers. To this aim, he proposed the following alternative :

$$R_{n,\alpha}(y) := \sum_{i=1}^{n}(\psi(\alpha(X_i - y))) = 0$$

with $\psi : \mathbb{R} \to \mathbb{R}$ an antisymmetric increasing function and $\alpha \in \mathbb{R}$ a parameter. One interesting choice of ψ is the following :

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2) & \text{if} \quad x \geq 0 \\ -\log(1 - x + x^2/2) & \text{if} \quad x < 0 \end{cases}$$

which gives the following graph :

`pdf`
  2

We thus have the following definition :

Definition : Catoni's mean estimator $\hat{\mu}_{\alpha,n}$ is the unique value y such that $R_{n,\alpha}(y) = 0$.

With some calculations, we can find that Catoni's mean estimator has a sub-Gaussian performance under some conditions.

Theorem : Let $X_1, ..., X_n$ be i.i.d. random variables with mean μ and variance σ^2. Let $\delta \in (0, 1)$ be such that $n > 2log(1 + \delta)$. Catoni's mean estimator $\hat{\mu}_{n,\alpha}$ with parameter

$$\alpha = \sqrt{\frac{2\log(1/\delta)}{n\sigma^2(1+\frac{2\log(1/\delta)}{n-2\log(1/\delta)})}}$$

satisfies

$$\mathbb{P}\left(|\hat{\mu}_{n,\alpha} - \mu| < \sqrt{\frac{2\sigma^2\log(1/\delta)}{n-2\log(1/\delta)}}\right) \geq 1-\delta$$

One disadvantage of Catoni's mean estimator compared to the median-of-means estimator is that Catoni's estimator depends on the variance σ^2. Furthermore, Catoni's mean estimator has the same problem than the median-of-means estimator in the sense that it depends on the confidence parameter δ and we cannot get rid of this dependence without becoming non-sub-Gaussian.

2.2 Multiple dimensions case

We want to develop the framework that we introduced before in the multiple dimensions case. Let $X_1, ..., X_n$ be i.i.d. random vectors with mean vector μ and covariance matrix $\Sigma = \mathbb{E}(X_i - \mu)(X_i - \mu)^T$. Again, the simplest choice for an estimator of the mean is the sample mean $\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ but it has the same problem as the sample mean estimator in the uni-dimensional case in the sense that when the distribution is heavy-tailed, it has a sub-optimal performance.

We also need to define what is a sub-Gaussian performance in the multi-dimensional case. As before, we want our estimator to be "close" to the true mean with high probability. The first problem that arises is how to define "close" in multiple dimensions. One natural way to measure distance in the space is the Euclidean norm. We will stick to this norm in this chapter.

Definition : A vector mean estimator is sub-Gaussian if it satisfies an inequality of the form

$$\mathbb{P}\left(||\overline{\mu}_n - \mu|| \leq \sqrt{\frac{Tr(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{max}\log(1/\delta)}{n}}\right) \geq 1-\delta$$

where $\lambda_{max} = \lambda_{max}(\Sigma)$ is the largest eigenvalue of the covariance matrix Σ.

We recall that the mean-squared error of the empirical mean equals

$$\mathbb{E}||\overline{\mu}_n - \mu||^2 = \frac{Tr(\Sigma)}{n} \iff \mathbb{E}||\overline{\mu}_n - \mu|| \leq \sqrt{\frac{Tr(\Sigma)}{n}}$$

One important remark to make is that the random fluctuations in the sub-Gaussian definition are controlled by λ_{max} which can be much smaller than $Tr(\Sigma)$.

2.2.1 Multivariate median-of-means

Now that we have defined the sub-Gaussianity in the multivariate case, we want to find robust estimators as in the uni-dimensional case. The first one that we studied before was the median-of-means estimator. The first problem when we try to define the median-of-means in the multivariate case is that we don't have a standard notion of median. We could use the coordinate-wise median, the geometric median, the Tukey median, the Oja median or the Liu median.

Suppose that we choose one of the mentioned median. Let's define the median-of-means estimator in the multivariate case.

Definition : Let $1 \leq k \leq n$ and partition $[n] = \{1, ..., n\}$ into k blocks $B_1, ..., B_k$ each of size $|B_i| \geq \lfloor n/k \rfloor \geq 2$. Given $X_1, ..., X_n$, compute the sample mean in each block $Z_j := \frac{1}{B_j} \sum_{i \in B_j} X_i$ and define the median-of-means estimator by $\hat{\mu}_n = M(Z_1, ..., Z_k)$ where $M(X_1, ..., X_n)$ is the median that we chose.

By using the coordinate-wise median, we don't find any interesting result in the sense that the estimator is far from sub-Gaussian. But we can try another approach. Let's define a non-standard notion of median.

Definition : Consider $(Z_1, ..., Z_k)$ the sample means in each block as defined in the definition of the median-of-means. We choose $\hat{\mu}_n$ to be the point in $\mathbb{R}^d$ with the property that the Euclidean ball centered at $\hat{\mu}_n$ that contains more than $k/2$ of the points Z_j has minimal radius.

With this notion of median, we find the following result :

Proposition : Let $X_1, ..., X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. Let $\delta \in (0, 1)$ and let $\hat{\mu}_n$ be the estimator defined above with $k = \lceil 8 \log(1/\delta) \rceil$. Then we have

$$\mathbb{P}\left(||\hat{\mu}_n - \mu|| \leq 4\sqrt{\frac{Tr(\Sigma)(8\log(1/\delta) + 1)}{n}} \right) \geq 1 - \delta$$

This result is very interesting as it is dimension-free and we only assume the existence of the covariance matrix. However, this is not a sub-Gaussian estimator and the notion of median used is very problematic from a computational point of view.

Suppose now that we decide to work with the geometric median. First, let's define what the geometric median is.

Defintion : Consider $(Z_1, ..., Z_k)$ the sample means in each block as defined in the definition of the median-of-means. The geometric median is defined as

$$\hat{\mu}_n = \underset{m \in \mathbb{R}^d}{\arg\min} \sum_{j=1}^{k} ||Z_i - m||$$

We can show that the geometric median-of-means estimator achieves a bound close to the one in the above proposition. This is still not a sub-Gaussian estimator but there exists efficient ways of computing the geometric median which is an important advantage.

In order to have estimators that are really sub-Gaussian, we need to define new estimators.

2.2.2 Median-of-means tournaments

Again, consider an i.i.d. sample $X_1, ..., X_n$ of random vectors in $\mathbb{R}^d$. We partition the set $\{1, ...n\}$ into k blocks $B_1, ..., B_k$, each of size $|B_j| \geq m := \lfloor n/k \rfloor$, where k is a parameter of the estimator whose value depends on the desired confidence level δ. For simplicity, let's assume that $k|n$ which implies that $|B_i| = m$ for all $i = 1, ..., k$. We define the sample mean within each block by $Z_j = \frac{1}{m}\sum_{i\in B_j} X_i$. For each $a \in \mathbb{R}^d$, let

$$T_a = \{x \in \mathbb{R}^d : \exists J \subset [k] : |J| \geq k/2 \text{ such that for all } j \in J, ||Z_j - x|| \leq ||Z_j - a||\}$$

and define the mean estimator by

$$\hat{\mu}_n \in \arg\min_{a\in\mathbb{R}^d} \text{radius}(T_a)$$

where $\text{radius}(T_a) = \sup_{x\in T_a} ||x - a||$. Thus, $\hat{\mu}_n$ is chosen to minimize, over all $a \in \mathbb{R}^d$, the radius of the set T_a defined as the set of points $x \in \mathbb{R}^d$ for which $||Z_j - x|| \leq ||Z_j - a||$ for the majority of the blocks.

T_a can be seen as the set of points in $\mathbb{R}^d$ that are at least as close to the point cloud $\{Z_1, ..., Z_k\}$ as the point a.

We have the two following results using the mean estimator defined above.

Theorem : Let $\delta \in (0, 1)$ and consider the mean estimator $\hat{\mu}_n$ with parameter $k = \lceil 200\log(2/\delta)\rceil$. If $X_1, ..., X_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ, then for all n we have

$$\mathbb{P}\left(||\hat{\mu}_n - \mu|| \leq \max\left(960\sqrt{\frac{Tr(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right)\right) \geq 1 - \delta$$

Theorem : Using the same notation as above and setting

$$r = \max\left(960\sqrt{\frac{Tr(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right)$$

with probability at least $1 - \delta$, for any $a \in \mathbb{R}^d$ such that $||a - \mu|| \geq r$, one has $||Z_j - a|| > ||Z_j - \mu||$ for more than $k/2$ indices j. In other words, $||a - \mu|| \geq r$ implies that $a \notin T_\mu$.

To prove these theorems, we use the notion of tournament (hence de name of the chapter). Let's define a tournament.

Definition : Let $\{1, ..., n\}$ be partitioned into k disjoint blocks $B_1, ..., B_k$ of size $m = n/k$. For $a, b \in \mathbb{R}^d$, we say that a defeats b (in a tournament) if

$$\frac{1}{m}\sum_{i\in B_j}(||X_i - b||^2 - ||X_i - a||^2) > 0$$

on more than $k/2$ blocks B_j.

2.3 Computational considerations

One thing that we need to keep in mind is that an estimator has to be usable in practice and not just a theoretical concept. An estimator is usable in practice if it can be computed in polynomial time. For example, the geometric median-of-means estimator is efficiently computable. On the other hand, the median-of-means tournament estimator is sub-Gaussian but not computable in polynomial time which is a huge issue.

Catoni's estimator for the mean is highly regarded for its computation-friendly nature, making it a practical and efficient choice for mean estimation. One of the key advantages of Catoni's estimator is its simplicity in terms of computational requirements. Unlike some other estimators that involve complex optimization algorithms or heavy computational procedures, Catoni's estimator can be easily computed using straightforward calculations. This simplicity not only reduces the computational burden but also enables faster and more efficient estimation, particularly when dealing with large datasets or real-time applications. We will keep this in mind for the following of our work.

But before applying all these techniques to SPD matrices, we make a small introduction on manifolds and we recall some properties of manifolds that will be useful for us.

3 Introduction on manifolds

3.1 Basic definitions and properties

Roughly speaking, a manifold is a topological space that locally resembles Euclidean space near each point. There are a lot of different manifolds, but the ones that will interest us are the Riemannian manifolds. These manifolds are of main interest because they include a geometric as well as a metric structure which allow us to compute things almost as in Euclidean spaces. Without going into the details, we define here the notions needed in order to characterize a Riemannian manifold.

We begin with the definition of a particular class of manifolds which are the most intuitive ones.

Definition : Let $F : \mathbb{R}^k \to \mathbb{R}^m$ be a differentiable map such that the Jacobian matrix $dF(x) = (\frac{\partial}{\partial x_j} F_i(x))^i_j$ has constant rank $k - d$ for all $x \in F^{-1}(0)$. Then the zero-level set $M = F^{-1}(0)$ is an embedded manifold of dimension d.

The map F gives an explicit representation of the manifold. This statement is a consequence of a well-known theorem in differential topology but we will consider it as a definition.

Now that we have a better idea of what a manifold could look like, let's define properly what a manifold is.

Definition : Let $x, y \in X$ be two points in a topological space X. We say that x and y are separated by neighborhoods if there exists a neighborhood U of x and a neighborhood V of y such that $U \cap V = \emptyset$.

Definition : A topological space X is a Hausdorff space if any two distinct points in X are separated by neighborhoods.

Definition : A topological space X is called locally Euclidean if there is a non-negative integer n such that every point in X has a neighborhood which is homeomorphic to real n-space $\mathbb{R}^n$.

Definition : A topological manifold is a locally Euclidean Hausdorff space.

Definition : A differentiable manifold (or smooth manifold) is a topological manifold with a globally defined differential structure.

Roughly speaking, a differentiable manifold is a topological manifold which is similar enough to a vector space to apply calculus on it. As it is similar to a vector space, we can generalize the notion of vector to differentiable manifolds :

Definition : The tangent space T_xM at point x of a differentiable manifold M is the set of all possible tangent vectors at x.

Intuitively, T_xM is a real vector space that contains the possible directions in which one can tangentially pass through x.

Definition : A Riemannian metric is a smoothly varying collection of scalar products $\langle \cdot, \cdot \rangle_x$ on each tangent space T_xM at points x of a manifold M. For each x, each such scalar product is a positive definite bilinear map $\langle \cdot, \cdot \rangle_x : T_xM \times T_xM \to \mathbb{R}$. The inner product gives a norm $|| \cdot || : T_xM \to \mathbb{R}$ by $||v||_x = \langle v, v \rangle_x$.

Definition : A Riemannian manifold is a differentiable manifold equipped with a Riemannian metric.

We say that the Riemannian metric is the intrinsic way of measuring length on a manifold. The extrinsic way is to consider the manifold as embedded in $\mathbb{R}^k$ and compute the length of a curve in M as for any curve in $\mathbb{R}^k$.

We introduce also the notion of complete manifold as it will be useful for us later.

Intuitive definition : A complete manifold is a Riemannian manifold for which, starting at any point p, one can follow a "straight" line indefinitely along any direction.

One interesting feature of Riemannian manifold is the curvature of the manifold. Intuitively, the curvature of a manifold describes how it is locally curved. For example, we know that a plane in $\mathbb{R}^3$ is flat whereas the sphere $\mathbb{S}^2$ is curved. The curvature gives us a mathematical interpretation of these intuitions.

There exists various ways of defining the curvature. For our purpose, we will use the notion of sectional curvature.

Definition : Let M be a Riemannian manifold and u, v two linearly independent tangent vectors at the same point of the manifold. The sectional curvature is given by

$$K(u, v) = \frac{\langle R(u, v)v, u \rangle}{\langle u, u \rangle \langle v, v \rangle - \langle u, v \rangle^2}$$

where $R(u, v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w$ is the Riemann curvature tensor and $\nabla_x y$ corresponds to the derivatives of a vector field y in the direction of another vector field x (cf. Levi-Civita connection for more details[7]).

First notice that the curvature depends on the choice of the Riemannian metric that we associate with the differentiable manifold. This means that we will have to be careful about the choice that we make for the metric when studying specific Riemannian manifolds.

Furthermore, it can be shown that the Riemann curvature tensor is a mapping from tangent vectors to tangent vectors, i.e.

$$\begin{aligned} R : T_xM \times T_xM \times T_xM &\longrightarrow T_xM \\ (u, v, w) &\longmapsto R(u, v)w \end{aligned}$$

This can be used to define another type of curvature, the Ricci curvature. First, we need to introduce the Ricci curvature tensor.

[7]Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression, Olivier Catoni, Ilaria Giulini, arXiv:1712.02747

Definition[8] **:** Let M be a Riemannian manifold. We define for each $p \in M$ the map $Ric_p : T_pM \times T_pM \to \mathbb{R}$ by $Ric_p(v,w) := Tr(u \mapsto R(u,v)w))$. Then the Ricci curvature tensor at point $p \in M$ is $Ric_p(v,w)$.

This complicated definition means that when you have fixed v and w, take any basis $u_1, ..., u_n$ of the tangent space T_pM and define $Ric_p(v,w) := \sum c_{ii}$ where for any fixed i, the numbers $c_{i1}, c_{i2}, ..., c_{in}$ are the coordinates of $R(u_i, v)w$ relative to the basis $u_1, ..., u_n$.

One can show that this definition does not depend on the choice of basis $u_1, ..., u_n$ that we make. Furthermore, we can see that the Ricci curvature tensor is symmetric, i.e. $Ric_p(v,w) = Ric_p(w,v)$ which implies that the Ricci curvature tensor can be completely determined by knowing $Ric_p(v,v)$ for all vectors $v \in T_pM$ of unit length.

Definition : Let M be a Riemannian manifold. The Ricci curvature at point $p \in M$ is given by the function

$$\begin{aligned} RicCurv_p : T_u &\longrightarrow \mathbb{R} \\ v &\longmapsto Ric_p(v,v) \end{aligned}$$

where T_u is the subset of all the unit tangent vectors in T_pM.

Notation : When we denote a Riemannian manifold by (M,g) where g denotes the Riemannian metric associated with the differentiable manifold M, we write $RicCurv(g)$ to speak about the Ricci curvature on the whole manifold.

This second notion of curvature appears to be really useful for us. Indeed, an important theorem relates directly with the Ricci curvature.

Theorem (Bonnet-Myers)[9] **:** Let (M,g) be a complete Riemannian manifold of dimension n whose Ricci curvature satisfies

$$RicCurv(g) \geq (n-1)k$$

for some positive real number k. Then M is compact and its diameter can be bounded by

$$diam(M) := \sup_{p,q \in M} dist(p,q) \leq \frac{\pi}{\sqrt{k}}$$

One could ask why we need such a theorem in order to find a robust estimator of the Fréchet mean. The answer is that we will work on non-positive curvature Riemannian manifolds in order to be sure that our spaces are not compact.

3.2 Symmetric positive definite matrices

One particularly interesting space is the space of symmetric positive-definite matrices. Indeed, this space is a non-positive curvature space and is frequently used in medical image analysis for example to represent the different regions in the brain. The space of SPD matrices is a subset of symmetric matrices and it can even be seen as a convex half-cone in the space of symmetric matrices which means that it is a smooth manifold and moreover that convex operations (like the mean !) are stable.

[8]Differential geometry, Marco Gualtieri, http://www.math.toronto.edu/mgualt/courses/18-367/docs/DiffGeomNotes-9.pdf
[9]The theory of manifolds, Ana Rita, https://math.mit.edu/~arita/18.101/manifolds-4.pdf

In order to prove that the space of SPD matrices is a non-positive curvature space, we need to define a Riemannian metric on it. First, we can ask what a tangent vector look like in $SPD(n)$. Let $P \in SPD(n)$ and consider a curve in $SPD(n)$ passing through P defined by $\gamma(t) = P + Qt + O(t)$. Then the tangent vector of this curve at point P is simply given by Q which is a symmetric matrix of same dimension as P with no other restrictions. Recall that the dimension of a $n \times n$ symmetric positive definition matrix is $n(n+1)/2$ because of the restrictions needed to ensure the symmetry as well as the positive-definite property.

As we are considering matrices, we take the standard Frobenius scalar product on matrices as our Riemannian metric, i.e. $\langle W_1, W_2 \rangle = Tr(W_1^T W_2)$. Of course, we could take other Riemannian metrics, but we will stick to the Frobenius scalar product for our purpose.

Theorem : The space $SPD(n)$ of symmetric positive definite matrices of dimension $n \geq 2$ equipped with the standard Frobenius scalar product as Riemannian metric has non-positive Ricci curvature.

One can show that this theorem still holds when we consider other affine-invariant Riemannian metrics, i.e. Riemannian metrics that respect the condition $\langle W_1, W_2 \rangle = \langle AW_1 + b, AW_2 + b \rangle$ for all invertible matrices A.

Now that we have a better idea of what the space $SPD(n)$ look like, one question arises. How can we compute a mean estimator in a such space ? And more generally, what is the mean of a sample in Riemannian manifolds ?

We answer to these questions in the following chapter.

3.3 Lie groups

There exist other interesting manifolds to study. Lie groups are an example of manifolds with a lot of nice properties.

Definition : A real Lie group is a group that is also a finite-dimensional real smooth manifold, in which the group multiplication and the group inverse are smooth maps. These two requirements can be combined to the single following requirement :

The mapping

$$\alpha : G \times G \to G, \quad \alpha(x, y) = x^{-1}y$$

is a smooth mapping of the product manifold into G.

The major property of Lie groups is that instead of studying the group itself, we can study its local version, called its Lie algebra. Indeed, we can assign to every Lie group a Lie algebra and if two Lie groups are isomorphic, their associated Lie algebras will be isomorphic. We then need to define what is a Lie algebra and what is the Lie algebra associated with a Lie group.

Definition : A Lie algebra is a vector space $\mathfrak{g}$ over a field F associated with a binary operation $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ called the Lie bracket satisfying the following properties :

- Bilinearity, i.e. $[ax + by, z] = a[x, z] + b[y, z]$ and $[x, ay + bz] = a[x, y] + b[x, z]$ for all $a, b \in F$ and $x, y, z \in \mathfrak{g}$.
- Alternativity, i.e. $[x, x] = 0$ for all $x \in \mathfrak{g}$.

- The Jacobi identity, i.e. $[x,[y,z]]+[y,[x,z]]+[z,[x,y]]=0$ for all $x,y,z\in\mathfrak{g}$.

The bilinearity along with the alternativity can be used to show that a Lie algebra satisfies the anticommutativity property, i.e. $[x,y]=-[y,x]$ for all $x,y\in\mathfrak{g}$.

In order to define what is the Lie algebra associated with a Lie group, we need to understand what is a left-invariant vector field.

Definition : Let G be a Lie group and $g\in G$. Then the maps $L_g : G\to G$ defined by $L_g(h)=gh$ and $R_g : G\to G$ defined by $R_g(h)=hg$ are called left-translation and right-translation respectively.

Obviously, we have that $L_g^{-1}=L_{g^{-1}}$ and $R_g^{-1}=R_{g^{-1}}$ which proves that these maps are in fact diffeomorphisms. We can then define the differential of these two maps.

Definition : The differential of L_g at a point $h\in G$ is given by the following map :

$$\begin{aligned} dL_g(h) : T_hG &\longrightarrow T_{gh}G \\ t &\longmapsto gt \end{aligned}$$

In a similar way, the differential of R_g at a point $h\in G$ is given by the following map :

$$\begin{aligned} dR_g(h) : T_hG &\longrightarrow T_{hg}G \\ t &\longmapsto tg \end{aligned}$$

Since L_g and R_g are diffeomorphisms for all $g\in G$, their differentials are vector-space isomorphisms for every $h\in G$.

We also need to define what is a vector field. Intuitively, we can think of a vector field on a differentiable manifold G as a collection of tangent vectors $X_p\in T_pG$ for $p\in G$, depending smoothly on the basepoints $p\in G$[10]. We can formalize this intuition with the following definition.

Definition : Let G be an n-dimensional differentiable manifold. A vector field X on G is a function from G to $TG=\{T_pG : p\in G\}$ which assigns to each point $p\in G$ a vector $X(p)\in T_pG$[11]. Furthermore, we say that X is smooth if $\pi\circ X=Id$ where $\pi : TG\to G$ is the canonical projection, i.e. $\pi(X(p))=p$ for all $p\in G$.

With this in mind, we can now define what is a left-invariant vector field.

Definition : A vector field X on a differentiable manifold G is called left-invariant if for any $g\in G$

$$dL_gX(h)=dL_g(X(h))=X(gh)\quad \forall h\in G$$

Roughly speaking, it means that dL_g maps the tangent vector at h of X to the tangent vector at gh of X and not another tangent vector, for all $h\in G$.

This allows us to finally define what is the Lie algebra of a Lie group !

Definition : The Lie algebra $Lie(G)$ of a Lie group G is the Lie algebra containing all smooth left-invariant vector fields on G.

[10] Mean estimation and regression under heavy-tailed distributions — a survey, Gabor Lugosi and Shahar Mendelson, arXiv:1906.04280

[11] An Introduction to Manifolds, Loring W. Tu, Universitext, 2007.

This definition is pretty abstract, but we can identify the Lie algebra of a Lie group way more simply !

Proposition : The Lie algebra $Lie(G)$ of a Lie group G is isomorphic to the tangent space at the identity element of G, i.e. $Lie(G) \cong T_eG$.

Proof : We need to prove that we can identify all the smooth left-invariant vector fields on G with elements of T_eG. By using dL_g, we have that all vectors in T_eG define a smooth left-invariant vector field. Indeed, let $t \in T_eG$. Then we can define $X_t = \{dL_g(e)(t) : g \in G\} = \{gt \in T_gG : g \in G\}$. This gives a set of tangent vectors, one for each point $g \in G$ which is precisely the definition of a vector field on a manifold.

Furthermore, X_t is left-invariant for all $t \in T_eG$. We have that

$$dL_gX_t(h) = dL_g(ht) = ght = X_t(gh) \quad \forall g, h \in G$$

The smoothness is trivial. Thus, we can construct a left-invariant vector field from each element of T_eG.

We need now to verify that every smooth left-invariant vector field can be defined by an element of T_eG. Let X be a smooth left-invariant vector field. By definition, we know that $dL_gX(h) = X(gh)$ and $\pi(X(g)) = g$ for all $g, h \in G$. In particular, we have that $dL_gX(e) = X(ge) = X(g)$ for all $g \in G$. This means that we can characterize X just by using $X(e) \in T_eG$ and left-translations, i.e. $X = \{dL_gX(e) : g \in G\}$. Thus, every smooth left-invariant vector field can be characterize by an element of T_eG.

This allows us to conclude that $Lie(G) \cong T_eG$.

$\square$

One could ask why it is so interesting to identify the Lie algebra of a Lie group. In fact, by definition of a Lie algebra, it is a vector space over a field K. Thus, it is a huge advantage to be able to study a vector space instead of a smooth manifold which is also a group because we can make a lot of things on a vector space that are not necessarily feasible on a Lie group directly.

But in order to have the choice of working on the Lie group or its Lie algebra, we need to have a mapping between these two. The mapping from the Lie algebra to the Lie group is called the exponential map.

Definition : Let G be a Lie group and $Lie(G)$ its Lie algebra. The exponential map is given by

$$\begin{aligned} \exp : Lie(G) &\longrightarrow G \\ X &\longmapsto \gamma(1) \end{aligned}$$

where $\gamma : \mathbb{R} \to G$ is the unique one-parameter subgroup of G whose tangent vector at the identity is equal to X.

As we will mostly work on matrix Lie groups, we have the following alternative definition, which is much more comprehensible.

Definition : The exponential map for matrix Lie groups coincide with the matrix exponential and is given by

$$\exp(X) = \sum_{i=0}^{\infty} \frac{X^k}{k!} = I + X + \frac{1}{2}X^2 + \frac{1}{6}X^3 + ...$$

where I is the identity matrix. This means that for matrix Lie groups, the exponential map is just the matrix exponential restricted to the Lie algebra associated with the considered Lie group.

This identification will be very useful for us later as the space of SPD matrices is a Lie group with its Lie algebra being the space of symmetric matrices.

In order to prove that $SPD(n)$ is a Lie group, we need to put a group structure on it such that the group multiplication and the group inverse are smooth maps[12]. We then define

$$\begin{aligned} \otimes : SPD(n) \times SPD(n) &\longrightarrow SPD(n) \\ (M, N) &\longmapsto \exp(\log(M) + \log(N)) \end{aligned}$$

where $\exp(X) = \sum_{i=0}^{\infty} \frac{X^k}{k!}$ and $\log(X)$ is defined as the matrix $Y \in M_n(\mathbb{R})$ such that $X = \exp(Y)$ if it exists. The advantage of using $SPD(n)$ is that the logarithm of a matrix $X \in SPD(n)$ always exists and it is furthermore symmetric. Thus, our operator $\otimes$ is well defined on $SPD(n) \times SPD(n) \to SPD(n)$. Furthermore, it is clear that mapping $(X, Y) \mapsto X^{-1} \otimes Y$ is smooth which proves that the multiplication and the inversion are smooth maps on $SPD(n)$. Thus, we can say that $SPD(n)$ is a Lie group.

Now, we want to understand what is its associated Lie algebra. We know that the Lie algebra of a Lie group G corresponds to the tangent space of G at the identity. For $SPD(n)$, the tangent space at each point correspond exactly to the space of $n \times n$ symmetric matrices, i.e. $Symm(n)$. Thus, we have that $Lie(SPD(n)) \cong Symm(n)$.

Furthermore, one can show that the exponential map from $Symm(n)$ to $SPD(n)$ is a bijection which means that we have a one-to-one mapping between elements of the Lie group and elements of the Lie algebra ! This will be a very important result for us in the following of this work as the logarithm map from $SPD(n)$ to $Symm(n)$ is then just the inverse of the exponential map.

3.4 Mean definition and estimation on Riemannian manifolds

When we talk about the mean in a simple setting such as unidimensional Euclidean space, it is not difficult to represent what we are studying. But the situation is quite different when we dive into different spaces, for example Riemannian manifolds. We need to define the mean in these spaces.

General definition : Let Q be a probability distribution on some space X and M a manifold containing X. The mean of the space X is given by

$$\mu = \arg\min_{p \in M} \int_X L(p, x) Q(dx)$$

where $L(p, x)$ is a loss function.

This very general definition can of course be restricted to a discrete set of points sampled from Q. This gives us the first general definition of an estimator.

Definition : Consider the same framework as before and let $x_1, ..., x_n$ sampled from Q. The empirical risk estimator is given by

[12] Riemannian geometric statistics in medical image analysis, Xavier Pennec, Stefan Sommer, Tom Fletcher, Elsevier, 2020.

$$\hat{\mu} = \arg\min_{p \in M} \frac{1}{n} \sum_{i=1}^{n} L(p, x_i)$$

By choosing special loss functions L, we finally arrive at the main subject of our work.

Definition : Let (M, g) be a complete metric space and $x_1, ..., x_n$ points in M. For any $p \in M$, define the Fréchet variance as

$$\psi(p) = \sum_{i=1}^{n} g^2(p, x_i)$$

The Fréchet mean m is then defined as the global minimizer of the Fréchet variance, i.e.

$$m = \arg\min_{p \in M} \sum_{i=1}^{n} g^2(p, x_i)$$

Roughly speaking, the Fréchet mean gives a central tendency of a sample of points in a metric space.

With this definition, we can finally understand what is a mean in a Riemannian manifold. Indeed, a Riemmanian manifold is always equipped with a Riemannian metric which makes it a metric space.

Now that we have a better understanding of what a manifold is, we can seek for techniques for robust mean estimation on manifolds.

4 Robust mean estimation on manifolds

There are various estimators for robust mean estimation on manifolds but our aim is to seek for computation-friendly estimators. Lizhen[13] works on this idea by generalizing the idea of "median-of-means" estimation that we already mentionned.

4.1 Geometric median

Let's first describe how to obtain a median on a manifold. For this, we define the geometric median on a manifold.

Definition : Let (M, g) be a metric space. If it exists, the geometric median p^* of points $p_1, ..., p_m \in M$ minimizes the sum of distances to the points, i.e.

$$p^* = med(p_1, ..., p_m) = \arg\min_{p \in M} \frac{1}{m} \sum_{i=1}^{m} g(p, p_i)$$

The first problem that arises when considering this definition is that there are various ways to metrize a manifold. If we consider the extrinsic way of metrize the manifold, we say that p^* is the extrinsic geometric

[13]Exponential Concentration for Geometric-Median-of-Means in Non-Positive Curvature Spaces, Ho Yun, Byeong U. Park, arXiv:2211.17155

median whereas if we consider the intrinsic way of metrize the manifold, p^* is the intrinsic geometric median. With this in mind, we have the following lemma.

Lemma : Let $p_1, ..., p_m \in M$, $p^* = med(p_1, ..., p_m)$ as before. Then

- Let g be the extrinsic distance for some embedding $J : M \to \tilde{M} \subset \mathbb{R}^d$. Let $\omega \in M$, ψ be an angle between $J(\omega) - J(p^*)$ and the tangent space $T_{J(p^*)}\tilde{M}$ and let

$$C_\alpha = \frac{1-\alpha}{\sqrt{1-2\alpha}\cos(\psi) - \alpha \sin(\psi)}$$

  where $\alpha \in (0, \cot(\psi)\tan(\frac{\psi}{2}))$. If $g(\omega, p^*) \geq C_\alpha \epsilon$, then there exists an α portion of elements of $p_1, ..., p_m$ which are at least ϵ distance away from ω. That is, there exists an index set $T \subset \{1, ..., m\}$ with $|T| \geq \alpha m$, and $g(p_j, \omega) \geq \epsilon$ for any $j \in T$.

- Let g be an intrinsic distance on M with respect to some Riemannian structure. Let $\omega \in M$, the log map $\log_{p^*}$, namely, the inverse exponential map $\log_{p^*} = \exp_{p^*}^{-1}$ be K-Lipschitz continuous from $B(\omega, \epsilon)$ to $T_{p^*}M$ where the distance on $T_{p^*}M$ is the Euclidean distance, and let

$$C_\alpha = K(1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$$

  where $\alpha \in (0, \frac{1}{2})$. If $g(\omega, p^*) \geq C_\alpha \epsilon$, then there exists an α proportion of elements $p_1, ..., p_m$ which are at least ϵ distance away from ω.

There are various known manifolds with K-Lipschitz continuous log maps as required in the lemma above. This is the case for the following manifold.

Proposition : The manifold of positive definite $n \times n$ matrices $PD(n)$ has a 1-Lipschitz continuous inverse exponential map at any $p \in PD(n)$. For a given metric, we have the following exponential and logarithm mappings

$$\exp_p A = p^{1/2} \exp(p^{-1/2} A p^{-1/2}) p^{1/2}$$

$$\log_p q = p^{1/2} \exp(p^{-1/2} q p^{-1/2}) p^{1/2}$$

where

$$\exp(X) = I + \frac{X}{1!} + \frac{X^2}{2!} + ... + \frac{X^n}{n!} + ...$$

$$\log(x) = (x - I) - \frac{(x-I)^2}{2} + ... + (-1)^{n-1}\frac{(x-I)^n}{n} + ...$$

for any $A, X \in Sym(n)$ and any $p, q, x \in PD(n)$.

4.2 Median-of-means on manifolds

Let $x_1, ..., x_n$ a dataset which is divided into m subsets $U_1, ..., U_m$ each of roughly size $\lfloor n/m \rfloor$. Let $\mu_1, ..., \mu_m$ be the optimizers of the empirical risk function from each subset, that is

$$\mu_j = \underset{p \in M}{\arg\min} L_{|U_j|}(p, U_j) \text{ for } j = 1, ..., m$$

f Our estimator μ^* is the geometric median of subset optimizers, i.e.

$$\mu^* = \arg\min_{p \in M} \sum_{j=1}^{m} g(p, \mu_j)$$

With this estimator, we have the following robustness properties :

Theorem : Let $\mu_1, ..., \mu_m$ be a collection of independent estimators of the parameter μ, and let $\mu^* = med(\mu_1, ..., \mu_m)$ be the geometric median.

- Let g be the extrinsic distance on M for some embedding $J : M \to \tilde{M} \subset \mathbb{R}^d$. Assume for any $\omega \in M$ the angle between $J(\omega) - J(\mu^*)$ and the tangent space $T_{J(\mu^*)}\tilde{M}$ is no bigger than $\bar{\psi}$. For any $\alpha \in (0, \cot(\bar{\psi})\tan(\frac{\bar{\psi}}{2}))$, set
$$\bar{C}_\alpha = \frac{1-\alpha}{\sqrt{1-2\alpha}\cos(\bar{\psi}) - \alpha\sin(\bar{\psi})}$$
- Let g be an intrinsic distance on M with respect to some Riemannian structure. Assume $\log_{\mu^*}$ is K-Lipschitz continuous from $B(\mu^*, \epsilon)$ to $T_{\mu^*}M$. For any $\alpha \in (0, \frac{1}{2})$, set
$$\bar{C}_\alpha = K(1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$$

Under one of these two assumptions, if

$$P(g(\mu_j, \mu) > \epsilon) \leq \eta \text{ for } i = 1, ..., n$$

where $\eta < \alpha$ then
$$P(g(\mu^*, \mu) > \bar{C}_\alpha \epsilon) \leq \exp(-m\phi(\alpha, \eta))$$

where
$$\phi(\alpha, \eta) = (1-\alpha)\log\left(\frac{1-\alpha}{1-\eta}\right) + \alpha\log\left(\frac{\alpha}{\eta}\right)$$

One can remark that the choice of the number of subets m is of primary importance. Indeed, a large number of subset estimators will give more robustness and a tighter concentration around the true parameter. On the other hand, we must have enough data in each subset to ensure that each subset estimator behaves well and η is sufficiently small. For a given confidence interval ϵ, one can determine the number of subsets to achieve η and the desired bound on the concentration or confidence level.

With Lizhen's work, we can see that it seems reasonable to seek for a robust estimator on manifolds.

5 Catoni's estimator

We recall that the aim of this project is to find a robust mean estimator for SPD matrices which is computation-friendly. Through the first chapter, we saw that a very simple estimator was Catoni's estimator. Our aim is thus to generalize Catoni's estimator to multivariate Eulclidean spaces and then to symmetric positive definite matrices.

5.1 Catoni's function for multivariate Euclidean spaces

Catoni himself worked on this problem with the help of Giulini[14]. In their paper, they find a mean estimator for random vector as follows.

Let $X \in \mathbb{R}^d$ be a random vector and $X_1, ..., X_n$ be n independent samples of X. Denote by $\mathbb{S}_d = \{\theta \in \mathbb{R}^d : ||\theta|| = 1\}$ the unit sphere in $\mathbb{R}^d$ and I_d the identity matrix of size $d \times d$. Let $\rho_\alpha = N(\alpha, \beta^{-1} I_d)$ be the normal distribution of mean $\alpha \in \mathbb{R}^d$ and covariance matrix $\beta^{-1} I_d$ with $\beta \in \mathbb{R}_+^*$.

The idea of Catoni and Giulini is that, instead of directly trying to compute the mean vector $\mathbb{E}[X]$, they estimate each component of $\langle \alpha, \mathbb{E}[X] \rangle$ in each direction $\alpha \in \mathbb{S}_d$ of the unit sphere. They then define the estimator of $\langle \alpha, \mathbb{E}[X] \rangle$ as

$$\mathcal{E}(\alpha) = \frac{1}{n\lambda} \sum_{i=1}^{n} \int \psi(\lambda \langle \theta', X_i \rangle) d\rho_\alpha(\theta') = \frac{1}{n\lambda} \sum_{i=1}^{n} \mathbb{E}_{\theta'}[\psi(\lambda \langle \theta', X_i \rangle) \mid X_i] \quad \text{with } \alpha \in \mathbb{S}^d \text{ and } \lambda > 0$$

where ψ is given by

$$\psi(x) = \begin{cases} x - \frac{x^3}{6} & \text{if} \quad -\sqrt{2} \leq x \leq \sqrt{2} \\ \frac{2\sqrt{2}}{3} & \text{if} \quad x > \sqrt{2} \\ -\frac{2\sqrt{2}}{3} & \text{if} \quad x < -\sqrt{2} \end{cases}$$

This function is quite different than Catoni's function for the univariate case, but it has still some very nice properties. Indeed, like the previous Catoni's function, ψ behaves close to the identity near the origin and is such that $\exp(\psi(x))$ is bounded by polynomial functions $\forall x \in \mathbb{R}$. They prove it in the following lemma :

Lemma : For any $x \in \mathbb{R}$

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$

Since

- $AX + b \sim N(b, AA^T)$ when $X \sim N(0, I_d)$
- $X \sim N(\mu, \Sigma) \iff$ there exist $\mu \in \mathbb{R}^k$, $A \in \mathbb{R}^{k \times l}$ such that $X = AZ + \mu$ and $\forall n = 1, ..., l : Z_n \sim N(0, 1)$ i.i.d.
- If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

we can write

$$\theta' \sim N(\alpha, \beta^{-1} I_d) = \beta^{-1/2} N(0, I_d) + \alpha = \beta^{-1/2} \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} + \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \beta^{-1/2} Z_1 + \alpha_1 \\ \vdots \\ \beta^{-1/2} Z_n + \alpha_n \end{pmatrix}$$

with $Z_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$ $\forall 1 \leq i \leq n$ and thus, we have that

$$\lambda \langle \theta', X_i \rangle = \lambda(\beta^{-1/2} Z_1 + \alpha_1) \cdot X_{i1} + \lambda(\beta^{-1/2} Z_2 + \alpha_2) \cdot X_{i2} + ... + \lambda(\beta^{-1/2} Z_n + \alpha_n) \cdot X_{in}$$

[14] Geometric means in a novel vector space structure on symmetric positive-definite matrices, Vincent Arsigny, Pierre Fillard, Xavier Pennec, Nicholas Ayache, http://www-sop.inria.fr/asclepios/Publications/Arsigny/arsigny__siam__tensors.pdf

$$\iff \lambda\langle\theta', X_i\rangle \sim N\left(\lambda\alpha_1 X_{i1}, \lambda^2\beta^{-1}X_{i1}^2\right) + \ldots + N\left(\lambda\alpha_n X_{in}, \lambda^2\beta^{-1}X_{in}^2\right) = N\left(\lambda\sum_{j=1}^{n}\alpha_j \cdot X_{ij}, \lambda^2\beta^{-1}\sum_{j=1}^{n}X_{ij}^2\right)$$

$$\iff \lambda\langle\theta', X_i\rangle \sim N\left(\lambda\langle\alpha, X_i\rangle, \lambda^2\beta^{-1}||X_i||^2\right)$$

i.e. $\lambda\langle\theta', X_i\rangle$ follows a normal distribution with mean $\lambda\langle\alpha, X_i\rangle$ and variance $\lambda^2\beta^{-1}||X_i||^2$. Furthermore, the influence function ψ is piecewise polynomial which means that we can compute explicitly the estimator $\mathcal{E}(\alpha)$.

Lemma : Let $W \sim N(0,1)$ be a standard Gaussian real-valued random variable. For any $m \in \mathbb{R}$ and any $\sigma \in \mathbb{R}_+$, define

$$\phi(m,\sigma) = \mathbb{E}[\psi(m + \sigma W)]$$

Then the function ϕ can be explicitly computed as

$$\phi(m,\sigma) = m\left(1 - \frac{\sigma^2}{2}\right) - \frac{m^3}{6} + r(m,\sigma)$$

where the correction term $r(m,\sigma)$ is just a function of the cumulative distribution function F of W, namely

$$\begin{aligned}
r(m,\sigma) =& \frac{2\sqrt{2}}{3}\left[F\left(\frac{-\sqrt{2}+m}{\sigma}\right) - F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right] - \left(m - \frac{m^3}{6}\right)\left[F\left(\frac{-\sqrt{2}+m}{\sigma}\right) + F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right] \\
&+ \sigma\frac{2-m^2}{2\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right) - \exp\left(-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right)\right] \\
&+ \frac{m\sigma^2}{2}\left\{F\left(\frac{-\sqrt{2}+m}{\sigma}\right) + F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right\} \\
&+ \frac{m\sigma^2}{2\sqrt{2\pi}}\left\{\frac{\sqrt{2}+m}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right] + \frac{\sqrt{2}-m}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right]\right\} \\
&+ \frac{\sigma^3}{6\sqrt{2\pi}}\left\{\left[\left(\frac{\sqrt{2}-m}{\sigma}\right)^2 + 2\right]\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right] - \left[\left(\frac{\sqrt{2}+m}{\sigma}\right)^2 + 2\right]\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right]\right\}
\end{aligned}$$

With this lemma in mind, we can now find an explicit formula for the estimator $\mathcal{E}(\alpha)$:

$$\mathcal{E}(\alpha) = \frac{1}{n\lambda}\sum_{i=1}^{n}\mathbb{E}[\psi(\lambda\langle\theta', X_i\rangle)] = \frac{1}{n\lambda}\sum_{i=1}^{n}\phi(\lambda\langle\alpha, X_i\rangle, \lambda\beta^{-1/2}||X_i||)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\langle\alpha, X_i\rangle\left(1 - \frac{\lambda^2||X_i||^2}{2\beta}\right) - \frac{\lambda^2\langle\alpha, X_i\rangle^3}{6} + r(\lambda\langle\alpha, X_i\rangle, \lambda\beta^{-1/2}||X_i||)$$

The interesting feature about this estimator is the fact that it has almost a sub-Gaussian performance ! Indeed, we have the following proposition :

Proposition : Assume that

- $\mathbb{E}\left[||X||^2\right] = Tr\left(\mathbb{E}[XX^T]\right) \leq T < +\infty$
- $\sup_{\alpha\in S}\mathbb{E}\left(\langle\alpha, X\rangle^2\right) \leq v \leq T < +\infty$

where T and v are two known constants and $S \subset \mathbb{S}^d$ is an arbitrary symmetric subset of the unit sphere, i.e. if

$\alpha \in S$ then $-\alpha \in S$. Let $\delta \in]0,1[$ be a confidence parameter and set the constants λ and β in the definition of $\mathcal{E}$ to

$$\lambda = \sqrt{\frac{2\log(1/\delta)}{nv}} \quad \text{and} \quad \beta = \sqrt{nT}\lambda = \sqrt{\frac{2T\log(1/\delta)}{v}}$$

Then, we have that

$$\mathbb{P}\left(\sup_{\alpha\in S}|\mathcal{E}(\alpha) - \langle\alpha, \mathbb{E}[X]\rangle| \leq \sqrt{\frac{T}{n}} + \sqrt{\frac{2v\log(1/\delta)}{n}}\right) \geq 1-\delta$$

Furthermore, if $\hat{m} \in \mathbb{R}^d$ is an estimator of $\mathbb{E}[X]$ that satisfies

$$\sup_{\alpha\in S}|\mathcal{E}(\alpha) - \langle\alpha, \hat{m}\rangle| \leq \sqrt{\frac{T}{n}} + \sqrt{\frac{2v\log(1/\delta)}{n}}$$

Then such a vector exists and we have that

$$\mathbb{P}\left(\sup_{\alpha\in S}|\langle\alpha, \hat{m} - \mathbb{E}[X]\rangle| \leq \sup_{\alpha\in S}|\mathcal{E}(\alpha) - \langle\alpha, \hat{m}\rangle| + \sqrt{\frac{T}{n}} + \sqrt{\frac{2v\log(1/\delta)}{n}}\right) \geq 1-\delta$$

$$\iff \mathbb{P}\left(\sup_{\alpha\in S}|\langle\alpha, \hat{m} - \mathbb{E}[X]\rangle| \leq 2\sqrt{\frac{T}{n}} + 2\sqrt{\frac{2v\log(1/\delta)}{n}}\right) \geq 1-\delta$$

For the particular case of $S = \mathbb{S}^d$, we thus obtain the following bound :

$$\mathbb{P}\left(||\hat{m} - \mathbb{E}[X]|| = \sup_{\alpha\in\mathbb{S}^d}|\langle\alpha, \hat{m} - \mathbb{E}[X]\rangle| \leq 2\left(\sqrt{\frac{T}{n}} + \sqrt{\frac{2v\log(1/\delta)}{n}}\right)\right) \geq 1-\delta$$

Thus, if we manage to find a mean estimator that satisfies the required properties stated in the theorem, we could have a mean estimator with sub-Gaussian performance by taking $T = Tr\left(\mathbb{E}[XX^T]\right)$ and $v = \lambda_{max}$ with λ_{max} the largest eigenvalue of the matrix XX^T.

One question is still unanswered. How can we have obtain a mean estimator from $\mathcal{E}(\alpha)$? Indeed, $\mathcal{E}(\alpha)$ is a one dimension estimator but we want to have an estimator for the mean of a sample in multiple dimensions.

The mean estimator coming from $\mathcal{E}(\alpha)$ is defined as follows :

$$\hat{m}_{Catoni} = \arg\min_{m\in\mathbb{R}^d} \sum_{\theta_i}(\langle m, \theta_i\rangle - \mathcal{E}(\theta_i))^2$$

Now that we know that there exists a sort of Catoni's function for the multivariate Euclidean spaces, our aim is to transpose this work into the framework of manifolds and more particularly into SPD matrices.

Our aim is to generalize Catoni's estimator for multivariate Euclidean spaces to Lie groups. Recall that Catoni's estimator for multivariate Euclidean spaces was given by

$$\mathcal{E}(\alpha) = \frac{1}{n\lambda}\sum_{i=1}^{n}\int \psi(\lambda\langle\theta', X_i\rangle)d\rho_\alpha(\theta') = \frac{1}{n\lambda}\sum_{i=1}^{n}\mathbb{E}_{\theta'}[\psi(\lambda\langle\theta', X_i\rangle) \mid X_i] \quad \text{with } \alpha \in \mathbb{S}^d \text{ and } \lambda > 0$$

where, instead of directly computing an estimator for $\mathbb{E}(X)$, they compute an estimator for each component of $\langle\alpha, X\rangle$ in each direction $\alpha \in \mathbb{S}_d$ of the unit sphere.

5.2 Catoni's estimator for SPD matrices

Let $X_1, ..., X_n \in SPD(n)$ be a sample of symmetric positive definite matrices. Our aim is now to estimate the mean of this sample. Let's define our strategy.

1. We transform our sample $X_1, ...X_n$ of $SPD(n)$ into elements $Y_1, ..., Y_n$ of $Symm(n)$ by using the logarithm map from the Lie group to the Lie algebra.
2. We vectorize each element Y_i of our sample into an element z_i of $\mathbb{R}^d$ with $d = \frac{n(n+1)}{2}$.
3. We compute Catoni's estimator $\bar{z}$ with our transformed sample in $\mathbb{R}^d$ by taking the optimal values of λ and β in Chapter 4.
4. We "devectorize" our candidate $\bar{z}$ to have an element $\bar{Y} \in Symm(n)$.
5. We finally go back to $SPD(n)$ by applying the exponential map to $\bar{Y}$ to find $\bar{X} = \exp(\bar{Y}) \in SPD(n)$.

This strategy seems pretty good, but some points are pretty obscure for now. Indeed, the third as well as the fifth points imply things that we never defined. We then develop this notion of "vectorization".

5.3 Vectorization

Definition : Let $X \in M_n(\mathbb{R})$. The vectorization of X is a linear transformation which converts the matrix into a vector. More precisely, let x_{ij} represent the element in the i-th row, j-th column in X. Then

$$vec(X) = [x_{11}, x_{21}, ..., x_{n1}, x_{12}, ..., x_{n2}, ..., x_{1n}, ..., x_{nn}]^T$$

For $X \in Symm(n)$, we know that there are only $\frac{n(n+1)}{2} = d$ independent entries as the entries in the upper triangular part of the matrix define also the entries in the lower triangular part, which means that $vec(X)$ can be viewed as a vector in $\mathbb{R}^d$ with a small adaptation of the definition.

Definition : Let $X \in Symm(n)$. Let x_{ij} represent the element in the i-th row, j-th column in X. Then the vectorization of X is given by

$$vec(X) = [x_{11}, x_{12}, x_{22}, ..., x_{1n}, ..., x_{nn}]^T \quad \text{with} \quad X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1n} \\ x_{12} & x_{22} & ... & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{n2} & ... & x_{nn} \end{pmatrix}$$

Furthermore, we can go back to a symmetric matrix from the vectorization. Indeed, let $x = (x_1, ..., x_d)^T \in \mathbb{R}^d$ with $d = \frac{n(n+1)}{2}$. Then the matrix X associated with x is given by

$$X = \begin{pmatrix} x_1 & x_2 & x_4 & ... & x_{d-n} \\ x_2 & x_3 & x_5 & ... & x_{d-n+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{d-n} & x_{d-n+1} & x_{d-n+2} & ... & x_d \end{pmatrix}$$

With this new notion of vectorization, we are now able to perform some computation ! Indeed, all the steps in our strategy are well defined and we can now test if Catoni's estimator is a good mean estimator in the context of symmetric positive definite matrices.

5.4 Simulation study

Catoni's estimator is a good estimator when the true mean is around 0 but it is not a translation-invariant estimator which makes it bad when the true mean is far from 0. Thus, our aim is to find a procedure in order to have a translation-invariant estimator.

Our strategy is the following : we observe a sample $X_1, ..., X_n$ of $n \times n$ SPD matrices. We vectorize them to have a sample $x_1, ..., x_n \in \mathbb{R}^{n(n+1)/2}$. Then we compute the median m_i component by component of our new sample to find the median $m \in \mathbb{R}^{n(n+1)/2}$. We subtract the median from $x_1, ..., x_n$ to get a sample $y_1, ..., y_n$ which is around 0. After that, we compute Catoni's estimator $\hat{m}_{Catoni}$ of $y_1, ...y_n$ and we add the median m to it in order to have a mean estimator of $x_1, ..., x_n$. After that, we can apply the strategy defined above to find a mean estimator which is an SPD matrix.

In order to have a good Catoni's estimator, it is important that the sample size is not too big. Indeed, if this is not the case, we know that the empirical mean will have a very good performance by the central limit theorem. Thus, we restrict ourselves to samples of size between 50 and 500.

We also need a tool to compare the performance of Catoni's estimator and the empirical mean. For this, we use the following definition of residual :

$$r = \sum_{i=1}^{n} |\hat{m}_i - \mu_i|$$

where $\hat{m}$ is a mean estimator and μ is the true mean before the "devectorization" step. We will consider that Catoni's estimator is better than the empirical mean when $r_{Catoni} < r_{emp}$.

For our simluation study, we produce 100 datasets of 3x3 SPD matrices where each element of the matrix comes from the following heavytailed distribution :

$$x_{i,j} \sim s \cdot Pareto(x_m, \alpha) \cdot Ber(p) + t := Heavytail(s, x_m, \alpha, p, t)$$

where

$$s \sim Unif(0, 5)$$

$$x_m \sim Unif(0, 10)$$

$$\alpha \sim Unif(1, 2)$$

$$t \sim Unif(0, 15)$$

and p is fixed at 0.5.

In fact, we produce directly 6 dimensions vectors for our dataset as the logarithm and the exponential map are bijective.

By applying the empirical mean and Catoni's estimator for each dataset, we get the following results for $r_{Catoni} - r_{emp}$:

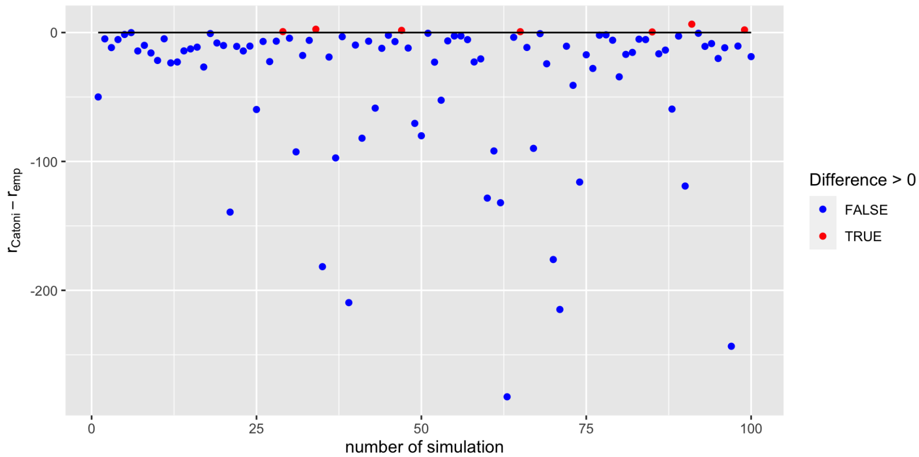

Figure 1: Difference between r_{Catoni} and r_{emp}

We can see that Catoni's estimator is 93 times better than the empirical mean out of 100 simulations !

We can also see this improvement by plotting the ellipsoid associated with the eigenvectors of the SPD matrices. Indeed, with a 3x3 SPD matrix, we can plot the ellipsoid which has $\lambda_i v_i$ with $i = 1, 2, 3$ as semi-axes where λ_i is the i-th eigenvalue and v_i is the i-th eigenvector of the SPD matrix. Thus, by plotting three ellipsoids associated with the true mean, the empirical mean and Catoni's estimator respectively, we can see if this improvement is really visible.

In order to have nice plots, we make 50 simulations coming from $Heavytail(s, x_m, \alpha, p = 0.5, t = 0)$ with $s \sim Unif(0, 0.5)$, $x_m \sim Unif(0, 10)$ and $\alpha \sim Unif(1, 2)$. In this way, we know that the true mean is just the identity matrix which means that the ellipsoid associated with the true mean will be the unit sphere. Thus, we will be able to compare the ellipsoids more easily.

Again, by plotting the residuals, we can see that Catoni's estimator is better than the empirical mean :

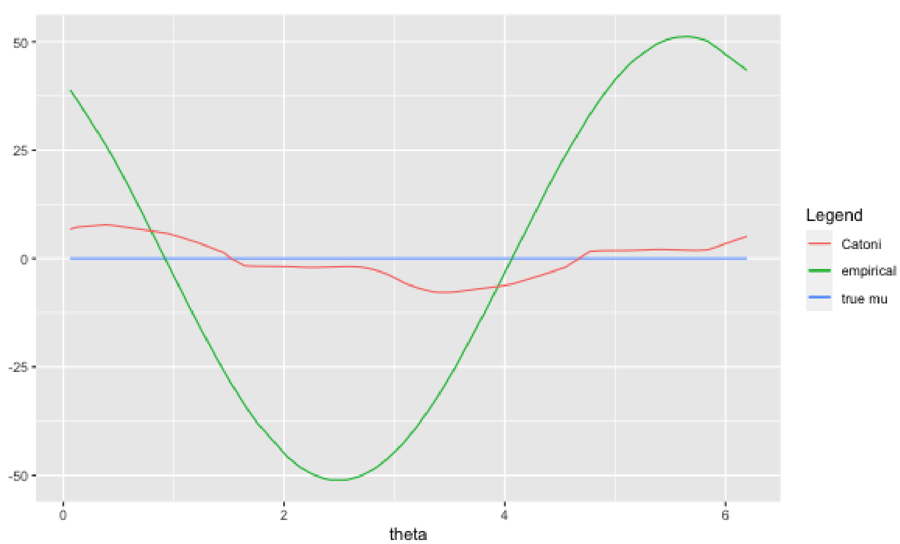

Figure 2: Difference between r_{Catoni} and r_{emp}

We now plot the ellipsoids for the third, ninth and tenth simulations. This will show us the three typical situations.

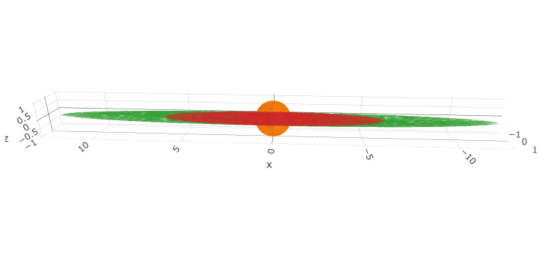

Figure 3: Ellipsoids for the third simulation. Orange is for the true mean, red is for Catoni's estimator and green is for the empirical mean.

For the third simulation, we see that Catoni's ellipsoid is closer to the true ellipsoid than the empirical ellipsoid. This means that Catoni's estimator estimates better the true mean than the empirical mean.

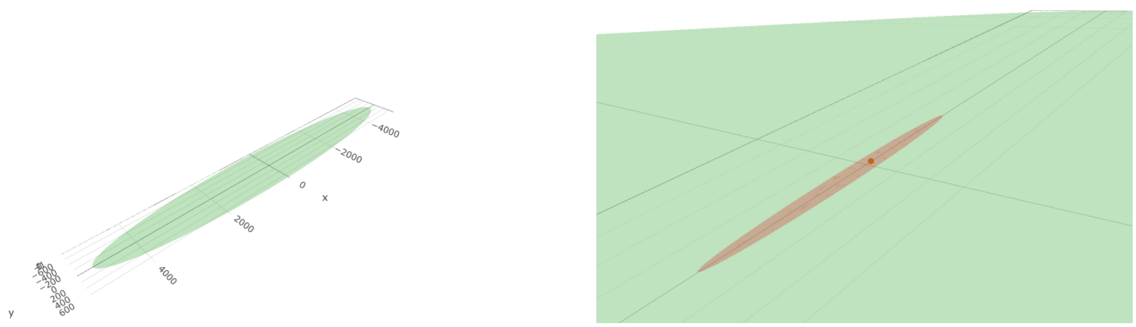

Figure 4: Ellipsoids for the ninth simulation. Orange is for the true mean, red is for Catoni's estimator and green is for the empirical mean.

For the ninth simulation, the empirical ellipsoid is so bad that we need to zoom in the figure to see appear the two other ellipsoids. Thus, we see here that the empirical mean is way worse than Catoni's estimator.

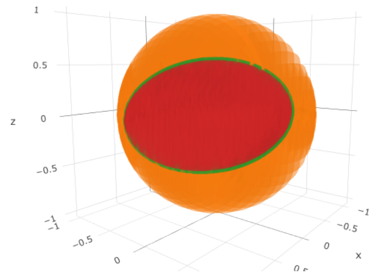

Figure 5: Ellipsoids for the tenth simulation. Orange is for the true mean, red is for Catoni's estimator and green is for the empirical mean.

The tenth simulation is one of the few simulations where the empirical mean is better than Catoni's estimator. This can be seen in the figure above. Indeed, Catoni's ellipsoid is inside the empirical ellipsoid which is itself inside the true ellipsoid. Thus, the empirical mean is a little bit better than Catoni's estimator for this simulation.

Thus, we see that our estimator seems in general better than the empirical mean when we use heavy-tailed samples.

5.5 Concentration of the estimators compared to the true mean

Remember the first chapter of this project where we introduced concentration properties of mean estimator. Our aim was to find estimators $\hat{\mu}$ such that the probability

$$\mathbb{P}(||\hat{\mu} - \mu|| > \delta)$$

decreased faster than the empirical mean when δ increased. Let's see if Catoni's estimator has a better performance than the empirical mean for our simulation study. To this aim, we take the Frobenius norm which is defined by

$$||X||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^2}$$

for all matrices $X \in M_n(\mathbb{R})$.

As we don't have the tools to compute the true probability, we use the empirical probability, i.e.

$$\hat{\mathbb{P}}(||\hat{\mu} - \mu||_F > \delta) = \frac{1}{N}\sum_{i=1}^{N}\{\hat{\mu} : ||\hat{\mu} - \mu||_F > \delta\}$$

where N is the number of simulations. As $||\hat{\mu} - \mu||_F$ are in general quite large, we will use the following concentration probability :

$$\hat{\mathbb{P}}(\log(||\hat{\mu} - \mu||_F) > \delta) = \frac{1}{N} \sum_{i=1}^{N} \{\hat{\mu} : \log(||\hat{\mu} - \mu||_F) > \delta\}$$

With this empirical probability, we get the following results for our two simulation studies from above :

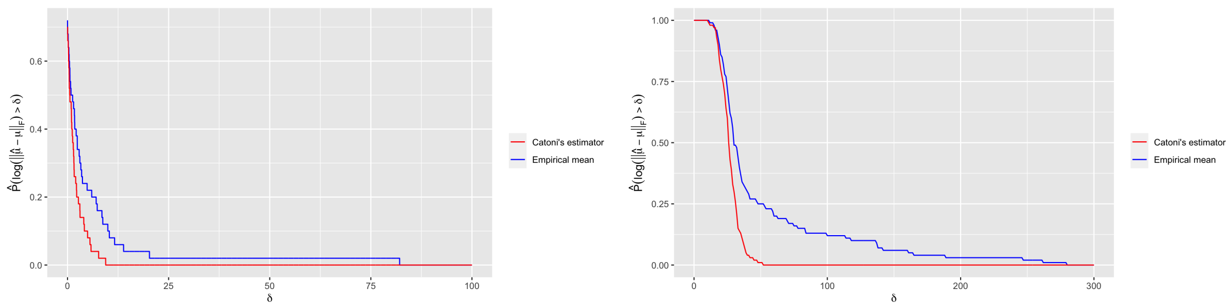

Figure 6: Concentration bound for Catoni's estimator (red) compared to the empirical mean (blue). Left is for the untranslated dataset, right is for the translated dataset.

Thus, we see that Catoni's estimator has an impressively better performance than the empirical mean.

6 Conclusion

In conclusion, this project presented a comprehensive analysis of Catoni's estimator on manifolds, supported by a simulation study on the specific space of 3×3 symmetric positive definite matrices. By extending the traditional estimator to the realm of manifolds, the research explored the potential of this method in handling high-dimensional data and capturing complex structures.

The simulation study demonstrated the effectiveness of Catoni's estimator on SPD matrices in various scenarios. The results consistently showcased its superior performance compared to the empirical mean, particularly when dealing with heavy-tailed distributions. The estimator exhibited remarkable robustness against this sort of distributions, highlighting its resilience in practical applications.

Moreover, the research revealed the versatility of Catoni's estimator on manifolds by demonstrating its applicability in diverse domains, such as computer vision, medical imaging, and sensor networks. This broadens the scope of its potential applications and emphasizes its relevance in solving real-world problems.

Overall, this project contributes to the existing literature by presenting a comprehensive analysis of Catoni's estimator on manifolds, combining theoretical insights with empirical evidence from a simulation study. The findings confirm the efficacy of this method in handling high-dimensional data coming from heavy-tailed distributions, and its potential for practical implementation. As the field of manifold estimation continues to evolve, further exploration of Catoni's estimator and its extensions promises to open up new avenues for statistical modeling and data analysis in manifold settings.

7 References

[1] Mean estimation and regression under heavy-tailed distributions — a survey, Gabor Lugosi and Shahar Mendelson, arXiv:1906.04280

[2] Robust optimization and inference on manifolds, Lizhen Lin, Drew Lazar, Bayan Saparpabayeva and David Dunson, arXiv:2006.06843

[3] Exponential Concentration for Geometric-Median-of-Means in Non-Positive Curvature Spaces, Ho Yun, Byeong U. Park, arXiv:2211.17155

[4] LECTURE 19: THE THEOREMS OF BONNET-MYERS, SYNGE AND PREISSMAN, http://staff.ustc.edu.cn/~wangzuoq/Courses/16S-RiemGeom/Notes/Lec19.pdf

[5] Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression, Olivier Catoni, Ilaria Giulini, arXiv:1712.02747

[6] Differential geometry, Marco Gualtieri, http://www.math.toronto.edu/mgualt/courses/18-367/docs/DiffGeomNotes-9.pdf

[7] The theory of manifolds, Ana Rita, https://math.mit.edu/~arita/18.101/manifolds-4.pdf

[8] Geometric means in a novel vector space structure on symmetric positive-definite matrices, Vincent Arsigny, Pierre Fillard, Xavier Pennec, Nicholas Ayache, http://www-sop.inria.fr/asclepios/Publications/Arsigny/arsigny_siam_tensors.pdf

[9] An Introduction to Manifolds, Loring W. Tu, Universitext, 2007.

[10] Riemannian geometric statistics in medical image analysis, Xavier Pennec, Stefan Sommer, Tom Fletcher, Elsevier, 2020.

[11] Exponential Concentration for Geometric-Median-of-Means in Non-Positive Curvature Spaces, Ho Yun, Byeong U. Park, arXiv:2211.17155