Statistical Estimation of Fréchet Mean Sets

Adam Quinn Jaffe

(with: Steven Neil Evans, Moïse Blanchard)

I. Introduction

Observation.

If $y_1, \ldots y_n$ are any points in $\mathbb{R}^k$, then

$$\begin{array}{rl} \text{minimize} & \displaystyle \frac{1}{n}\sum_{i=1}^{n} \|x - y_i\|^2 \\ \text{over} & x \in \mathbb{R}^k \end{array}$$

has a unique solution given by

$$x_* = \frac{y_1 + \cdots + y_n}{n}.$$

Observation.

If μ is a nice probability distribution on $\mathbb{R}^k$, then

$$\begin{array}{rl} \text{minimize} & \displaystyle\int_{\mathbb{R}^k} \|x - y\|^2 \, d\mu(y) \\ \text{over} & x \in \mathbb{R}^k \end{array}$$

has a unique solution given by

$$x_* = \int_{\mathbb{R}^k} y \, d\mu(y).$$

The preceding optimization problems depend on the **metric** structure of $\mathbb{R}^k$ but not on its **vector space** structure. Hence, one can define "averages" in general metric spaces.

Definition Let (X, d) be a metric space, and let μ a nice probability distribution on X. Then set the *Fréchet mean of* μ to be

$$F(\mu) := \operatorname*{arg\,min}_{x \in X} \int_X d^2(x, y)\, d\mu(y)$$

which is the **set** of minimizers.

Restating the observation: In $\mathbb{R}^k$ with its usual metric, we have

$$F\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}\right) = \left\{\frac{1}{n}\sum_{i=1}^{n} y_i\right\} \text{ and } F(\mu) = \left\{\int_X y\, d\mu(y)\right\}.$$

Example

Write $\mathbb{S}^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$ for the circle of radius 1 in the plane, endowed with its **geodesic** metric.

Write $\mu = \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(0,-1)}$.

Then we have $F(\mu) = \{(1,0), (-1,0)\}$.

Notice that $\mathbb{S}^1$ consists of two arcs which are "isomoprhic" to the usual real interval $[-\pi, \pi]$.

Example

Consider $X = [-1, 0) \cup (0, 1]$ with the metric **inherited** from $\mathbb{R}$.

Write $\mu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$.

Then $F(\mu)$ is empty.

The Fréchet mean "wants to be" the point $0 \in \mathbb{R}$ but $0 \notin X$.

Example

For $m \geq 2$, consider $X = \{1, 2, \ldots, m\}$ with the **discrete** metric.

Write $\mu := \frac{1}{m} \sum_{i=1}^{m} \delta_i$ for the uniform measure.

Then $F(\mu) = X$.

Fréchet means are an important tool for many problems in geometry and topology:

- Cartan's fixed point theorem,
- analysis of CAT(0) spaces, and
- metric thickenings

Additionally, they have become an important tool for *non-Euclidean statistics* wherein one attempts to understand data that have inherent geometry. For example, many practitioners deal with data assumed to live in various geometric settings:

- a Riemannian manifold (computer vision, shape theory),
- a particular graph or a space of graphs (network analysis),
- certain space of trees (computational phylogenetics), and
- many non-linear spaces of matrices.

This talk will focus on the statistical setting. Suppose:

- (X, d) is a fixed known metric space,
- μ is a fixed but unknown Borel probability measure on X, and
- $Y_1, Y_2, \ldots$ are independent identically-distributed (IID) data points which are sampled from the distribution μ.

Question. Can we estimate $F(\mu)$, by only using the data $Y_1, Y_2, \ldots$?

In the Euclidean setting, this is one of the most classical and one of the most well-understood problems in statistics.

In the non-Euclidean case, not much is known; things are complicated by the fact that these are random sets.

If one assumes $\#F(\mu) = 1$ a priori, then this task again becomes easy. But what can we say in general?

It is also possible to consider the *Fréchet p-mean of μ* for $1 \leq p \leq \infty$ as

$$F_p(\mu) := \arg\min_{x \in X} \int_X d^p(x, y)\, d\mu(y).$$

The most important cases are the following, which correspond to well-studied Euclidean counterparts:

- $p = 1$ for medians
- $p = 2$ for means, and
- $p = \infty$ for circumcenters.

In this talk we'll focus exclusively on $F := F_2$, but the other cases can be handled with a little extra work.

I. Introduction

II. Empirical Fréchet Means

III. Relaxed Empirical Fréchet Means

IV. Extensions and Applications

II. Empirical Fréchet Means

In Euclidean space, the strong law of large numbers (SLLN) states the following: If μ is a Borel probability measure on $\mathbb{R}^k$ and if $Y_1, Y_2, \ldots$ are IID data points which are sampled from μ, then $\int_{\mathbb{R}^k} \|y\| \, d\mu(y) < \infty$ implies

$$\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} = \int_{\mathbb{R}^k} y \, d\mu(y)$$

with probability one.

In other words, the empirical mean is a consistent estimator of the population mean.

Is there some kind of SLLN Fréchet means? That is, do we hve

$$\lim_{n\to\infty} F(\bar{\mu}_n) = F(\mu)$$

with probability one, where $\bar{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$ denotes the empirical measure of the first n samples?

The difficulty is that these are random sets, so we need to be careful about the notion of convergence.

We want to find a strong enough topology such that this convergence is interesting but weak enough that it is true and can be proven in great generality!

For a metric space (X, d) write $\mathrm{K}(X)$ for the space of non-empty compact subsets of X.

For $K, K' \in \mathrm{K}(X)$, define

$$\vec{d}_{\mathrm{H}}(K, K') = \max_{x \in K} d(x, K') = \max_{x \in K} \min_{x' \in K'} d(x, x')$$

for the *one-sided Hausdorff distance from* K to K'.

Equivalently, $\vec{d}_{\mathrm{H}}(K, K')$ is the smallest radius $r \geq 0$ such that K is contained in the r-thickening of K'.

Under what conditions on (X, d) and μ do we have

$$\lim_{n \to \infty} \vec{d}_{\mathrm{H}} \left(F(\bar{\mu}_n), F(\mu) \right) = 0$$

with probability one?

A brief history:

- Ziezold (1977): $\#X < \infty$
- Sverdrup-Thygesen (1981): X compact
- Evans-Jaffe (2020): X Heine-Borel and $\int_X d^2(x, y) \, d\mu(y) < \infty$
- Schötz (2020): X Heine-Borel and $\int_X d(x, y) \, d\mu(y) < \infty$.

This notion of convergence is like a guarantee of "no false positives" in the sense that all elements of $F(\bar{\mu}_n)$ are guaranteed to be close to some element of $F(\mu)$.

We say in these cases that $F(\bar{\mu}_n)$ is $\vec{d}_{\mathrm{H}}$-*consistent*.

The idea of the proof is to try to show that the map F is just plainly continuous.

Write $\mathcal{P}_2(X)$ for the space of Borel probability measures μ on X with $\int_X d^2(x,y)\,d\mu(y) < \infty$ for some $x \in X$. Write τ_w^2 for the topology on $\mathcal{P}_2(X)$ generated by the 2-Wasserstein metric.

Write $\mathrm{K}(X)$ for the set of non-empty compact subsets of X, and write τ_H^+ for the topology generated by "balls" of the one-sided Hausdorff distance. (Note: τ_H^+ is not T_2.)

Lemma (Evans-Jaffe 2020)

The function $F : (\mathcal{P}_2(X), \tau_w^2) \to (K(X), \tau_H^+)$ *is continuous.*

Then, the SLLN follows by simply showing that we have $\bar{\mu}_n \to \mu$ in τ_w^2 with probability one, which is easy.

It may be more natural to consider the *Hausdorff metric* defined for $K, K' \in \mathrm{K}(X)$ via

$$d_{\mathrm{H}}(K, K') := \max\left\{\vec{d}_{\mathrm{H}}(K, K'), \vec{d}_{\mathrm{H}}(K', K)\right\}.$$

Equivalently, this is the smallest $r \geq 0$ such that K is contained in the r-thickening of K', and K' is contained in the r-thickening of K.

Then, the convergence

$$\lim_{n\to\infty} d_{\mathrm{H}}\left(F(\mu), F(\bar{\mu}_n)\right) = 0$$

with probability one can be interpretd as a guarantee of both "no false positives and no false negatives".

We say in these cases that $F(\bar{\mu}_n)$ is d_{H}-*consistent*.

Under what conditions on (X, d) and μ is it that $F(\bar{\mu}_n)$ is d_{H}-consistent? Unfortunately:

Theorem (Evans-Jaffe 2020)

If $\#X < \infty$, then $F(\bar{\mu}_n)$ is d_H-consistent if and only if "$\#F(\mu) = 1$".

In order to state the " " part precisely, we need to introduce some form of quotienting away trivialities in the metric measure space (X, d, μ).

Now we interpret these results in the context of our statistical problem:

If we are aiming to estimate $F(\mu)$, then the estimator $F(\bar{\mu}_n)$ is:

- always $\vec{d}_{\mathrm{H}}$-consistent.
- not always d_{H}-consistent.

The problem is that $F(\bar{\mu}_n)$ often misses points that should be in $F(\mu)$.

In a sense, the set $F(\bar{\mu}_n)$ is plainly "too small".

III. Relaxed Empirical Fréchet Means

Let's return to the statistical setting. If our goal is to estimate $F(\mu)$ using only the data $Y_1, Y_2, \ldots$, then there is no reason we are stuck with the estimator $F(\bar{\mu}_n)$. In fact, we just saw that this estimator is rather bad!

Instead, we should try to come up with another estimator $\hat{F}_n : X^n \to \mathrm{K}(X)$ which has better properties.

Is there some estimator $\hat{F}_n$ which is d_{H}-consistent, that is, which has

$$\lim_{n\to\infty} d_{\mathrm{H}}(\hat{F}_n(Y_1, \ldots, Y_n), F(\mu)) = 0$$

with probability one?

Since $F(\bar{\mu}_n)$ is just "too small", it makes sense to try to "enlarge" it.

Definition. Let (X, d) be a metric space, let μ a nice probability distribution on X, and let $\varepsilon \geq 0$. Then the *ε-relaxed Fréchet mean of μ* is the set

$$F(\mu, \varepsilon) := \left\{ x \in X : \int_X d^2(x, y)\, d\mu(y) \leq \inf_{x' \in X} \int_X d^2(x', y)\, d\mu(y) + \varepsilon \right\}.$$

Thus, $F(\mu) = F(\mu, 0)$.

The idea is to choose the relaxation parameter ε_n carefully so that $\hat{F}_n(Y_1, \ldots, Y_n) := F(\bar{\mu}_n, \varepsilon_n)$ is in fact d_{H}-consistent.

As we previously stated, Schötz (2020) proved that $F(\bar{\mu}_n)$ is $\vec{d}_{\mathrm{H}}$-consistent. In fact, he proved more:

Theorem (Schötz 2020)

Suppose that (X, d) is a Heine-Borel metric space and that $\int_X d(x, y)\, d\mu(y) < \infty$ for some $x \in X$. If $\varepsilon_n \to 0$, then $F(\bar{\mu}_n, \varepsilon_n)$ is $\vec{d}_H$-consistent.

In words, adding a vanishing amount of relaxation does not ruin the $\vec{d}_{\mathrm{H}}$-consistency.

Thus, if we choose ε_n to vanish sufficiently slowly, then we might be able to enlarge $F(\bar{\mu}_n)$ enough to get d_{H}-consistency,

Theorem (Blanchard-Jaffe 2022)

Suppose that (X, d) is a Heine-Borel-Dudley metric space and that $\int_X d(x, y)\, d\mu(y) < \infty$ for some $x \in X$, and consider the relaxation parameter

$$\varepsilon_n = c\sqrt{\frac{\log \log n}{n}}$$

for any $c > 0$. There exists a constant $c_* = c_*(X, d, \mu)$ such that:

- *If $c > c_*$, then $F(\bar{\mu}_n, \varepsilon_n)$ is d_H-consistent.*
- *If $c < c_*$, then $F(\bar{\mu}_n, \varepsilon_n)$ is not d_H-consistent.*

In words, the relaxation rate $\varepsilon_n = c_* n^{-1/2}(\log \log n)^{1/2}$ provides the exact cutoff between consistency and inconsistency.

It seems that this result shows that we should always use

$$\varepsilon_n = c_* \sqrt{\frac{\log \log n}{n}}$$

as the relaxation parameter in practice. However, c_* depends on the distribution μ, which is unknown to us.

To get around this, we need to introduce a "multi-step" procedure: First estimate c_*, then use this to choose the appropriate relaxation scale.

procedure RelaxedFrechetMean
  Input: data $Y_1, \ldots, Y_n \in X$ and relaxation scale $\varepsilon_n \geq 0$
  Output: subset $F(\bar{\mu}_n, \varepsilon_n) \subseteq X$ and optimal objective $m(\bar{\mu}_n) \geq 0$
end procedure

procedure TwoStepEstimator
  Input: data $Y_1, \ldots, Y_n \in X$
  Output: subset $F_n^{(2)} \subseteq X$
  ▷ step 0: no relaxation
  $(\_\_, c_n^{(0)}) \leftarrow$ RelaxedFrechetMean$(Y_1, \ldots, Y_n, 0)$
  ▷ step 1: consistent but sub-optimal relaxation
  $\varepsilon_n^{(1)} \leftarrow c_n^{(0)} n^{-1/2} (\log n)^{1/2}$
  $(F_n^{(1)}, \_\_) \leftarrow$ RelaxedFrechetMean$(Y_1, \ldots, Y_n, \varepsilon_n^{(1)})$
  $c_n^{(1)} \leftarrow$ **maximize** $\frac{1}{n} \sum_{i=1}^{n} (d^2(x, Y_i) - d^2(x', Y_i))^2$
    $-(\frac{1}{n} \sum_{i=1}^{n} (d^2(x, Y_i) - d^2(x', Y_i)))^2$
    over $x, x' \in F_n^{(1)}$
  ▷ step 2: near-optimal relaxation
  $\varepsilon_n^{(2)} \leftarrow \frac{3}{2}\, c_n^{(1)} n^{-1/2} (\log \log n)^{1/2}$
  $(F_n^{(2)}, \_\_) \leftarrow$ RelaxedFrechetMean$(Y_1, \ldots, Y_n, \varepsilon_n^{(2)})$
  return $F_n^{(2)}$
end procedure

IV. Extensions and Applications

- A key feature of the “topological proofs” of consistency is that they can be used to “push forward” other limit theorems for $\bar{\mu}_n$ down to limit theorems for $F(\bar{\mu}_n)$. (large deviations principle, ergodic theorem, etc.)
- The relaxation scales $\varepsilon_n^{(2)}$ from the algorithm are themselves random! So, we need to extend the Schötz (2020) SLLN to the case that $\varepsilon_n \to 0$ with random variables that are correlated with the data.
- Very similar methods can be used to prove strong consistency for k-means clustering and adaptive variants thereof. (Jaffe 2021)
- Similar methods can be used to study asymptotics of ill-posed M-estimation problems. (in progress)