Inconsistency of template estimation by minimizing of the variance/pre-variance in the quotient space

Loïc Devilliers, Stéphanie Allassonnière, Alain Trouvé, Xavier Pennec

Inconsistency of Template Estimation by Minimizing of the Variance/Pre-Variance in the Quotient Space †

Loïc Devilliers*, Stéphanie Allassonnière†, Alain Trouvé‡, and Xavier Pennec§

June 21, 2017

Abstract

We tackle the problem of template estimation when data have been randomly deformed under a group action in the presence of noise. In order to estimate the template, one often minimizes the variance when the influence of the transformations have been removed (computation of the Fréchet mean in the quotient space). The consistency bias is defined as the distance (possibly zero) between the orbit of the template and the orbit of one element which minimizes the variance. In the first part, we restrict ourselves to isometric group action, in this case the Hilbertian distance is invariant under the group action. We establish an asymptotic behavior of the consistency bias which is linear with respect to the noise level. As a result the inconsistency is unavoidable as soon as the noise is enough. In practice, template estimation with a finite sample is often done with an algorithm called "max-max". In the second part, also in the case of isometric group finite, we show the convergence of this algorithm to an empirical Karcher mean. Our numerical experiments show that the bias observed in practice can not be attributed to the small sample size or to a convergence problem but is indeed due to the previously studied inconsistency. In a third part, we also present some insights of the case of a non invariant distance with respect to the group action. We will see that the inconsistency still holds as soon as the noise level is large enough. Moreover we prove the inconsistency even when a regularization term is added.

*Université Côte d'Azur, Inria, France, loic.devilliers@inria.fr
†CMAP, Ecole polytechnique, CNRS, Université Paris-Saclay, 91128, Palaiseau, France
‡CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France
§Université Côte d'Azur, Inria, France

Contents

1 Introduction

1.1 General Introduction

Template estimation is a well known issue in different fields such as statistics on signals [KSW11], shape theory, computational anatomy [GMT00, JDJG04, CMT+04] etc. In these fields, the template (which can be viewed as the prototype of our data) can be (according to different vocabulary) shifted, transformed, wrapped or deformed due to different groups acting on data. Moreover, due to a limited precision in the measurement, the presence of noise is almost always unavoidable. These mixed effects on data lead us to study the consistency of algorithms which claim to compute the template. A popular algorithm consists in the minimization of the variance, in other words, the computation of the Fréchet mean in quotient space. This method has been already proved to be inconsistent [BC11, MHP16, DATP17]. In [BC11] the authors proves the inconsistency with a lower bound of the expectation of the error between the original template and the estimated template with a finite sample, they deduce that this expectation does not go to zero as the size of the sample goes to infinity. This work was done in a functional space, where functions only observed at a finite number of points of the functions were observed. In this case one can model these observable values on a grid. When the resolution of the grid goes to zero, one can show the consistency [PZ16] by using the Fréchet mean with the Wasserstein distance on the space of measures rather than in the space of functions. However, in (medical) images the number of pixels or voxels is finite.

In [MHP16], the authors demonstrated the inconsistency in a finite dimensional manifold with Gaussian noise, when the noisel level tends to zero. In our previous work [DATP17], we focused our study on the inconsistency with Hilbert Space (including infinite dimensional case) as ambient space. This current paper is an extension of a conference paper [DPA17].

1.2 Why Using a Group Action? Comparison with the Standard Norm

In the following, we take a simple example which justifies the use of the group action in order to compare the shape of two functions:

On Figure 1, suppose that you want to compare these functions. The simplest way to compare f_0 with f_1 would be to compute the L^2-norm (or any other norm) of $f_0 - f_1$, if we do that we have that $\|f_0 - f_1\| \simeq 0.6$. Likewise $\|f_0 - f_2\| \simeq 0.6$, therefore the norm tells us that f_0 is at the same distance from f_1 and from f_2. Yet, our eyes would say that f_0, f_1 have the same shape, contrarily to f_0 and f_2. Therefore the simple use of the L^2-norm in the space of functions is not enough. To have a relevant way to compare functions, one can register functions first. Firstly, we estimate the better time translation which aligns f_0 and f_1 and secondly, we compute the L^2-norm after this alignment step. On this example, we find that the distance is now $\simeq$0.02. On the contrarily, after alignment the distance between f_0 and f_2 is still $\simeq$0.6. With this new way of comparing

functions, the functions f_0 looks like f_1 but do not look like f_2. This fits with our intuition. That is why we use a group action in order to perform statistics. In the following paragraph, we precise how to do it in general.

This idea of using deformations/transformation in order to compare things is not new. It was already proposed by Darcy Thompson [Tho42] in the beginning of the 20th century, in order to classify species.

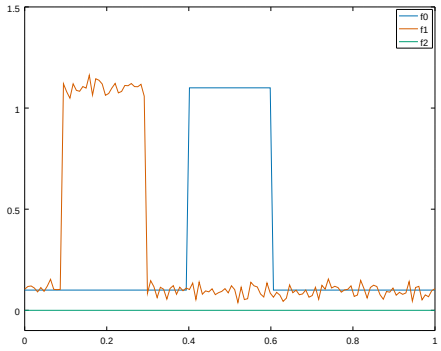


Figure 1: Three functions defined on the interval $[0, 1]$. The blue one (f_0) is a step function, the red one (f_1) is a translated version of the blue one when noise has been added, and the green one (f_3) is the null function.

1.3 Settings and Notation

In this paper, we suppose that observations belong to a Hilbert space $(M, \langle\cdot,\cdot\rangle)$, we denote by $\|\cdot\|$ the norm associated to the dot product $\langle\cdot,\cdot\rangle$. We also consider a group of transformation G which acts on M the space of observations. This means that $g' \cdot (g \cdot x) = (g'g) \cdot x$ and $e \cdot x = x$ for all $x \in M$, g, $g' \in G$, where e is the identity element of G. Note that in this article, $g \cdot x$ is the result of the action of g on x, and $\cdot$ should not to be confused with the multiplication of real numbers noted $\times$

The generative model is the following: we transform an unknown template $t_0 \in M$ with Φ a random and unknown element of the group G and we add some noise. Let σ be a positive noise level and ϵ a standardized noise: $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\|\epsilon\|^2) = 1$. Moreover we suppose that ϵ and Φ are independent random variables. Finally, the only observable random variable is:

$$Y = \Phi \cdot t_0 + \sigma\epsilon. \tag{1}$$

This generative model is commonly used in Computational anatomy in diverse frameworks, for instance with currents [DPC+14, GM01], varifolds [Cha13], LDDMM on images [BMTY05] but also in functional data analysis [KSW11]. All these works are applied in different spaces, for instance, the varifold builds an embedding of the surfaces into an Hilbert space, and a group of diffeomorphisms have the ability of deform these surfaces. Supposing a general group action on a space with the generative

model (1) allows us to embed all these various situations into one abstract model, and to study template estimation in this abstract model.

Example of noise: if we assume that the noise is independent and identically distributed on each pixel or voxel with a standard deviation w, then $\sigma = \sqrt{N}w$, where N is the number of pixels/voxels. However, the noise which we consider can be more general: we do not require the fact that the noise is independent over each region of the space M.

Note that the inconsistency of Template estimation can be also studied with an alternative generative model, called backward model where $Y = \Phi \cdot (t_0 + \sigma\epsilon)$ [DATP17]. Some authors also use the term *perturbation model* see [Huc11, Roh03, Goo91].

Quotient space: the random transformation of the template by the group leads us to project the observation Y into the quotient space. The quotient space is defined as the set containing all the orbit $[x] = \{g \cdot x,\, g \in G\}$ for $x \in M$. The set which is constituted of all orbits is call the quotient space M by the group G and is noted by:

$$Q = M/G = \{[x],\, x \in M\}.$$

As we want to do statistics on this space, we aim to equip the quotient with a metric. One often requires that d_M the distance in the ambient space is invariant under the group action G, this means that

$$\forall m, n \in M,\, \forall g \in G \quad d_M(g \cdot m, g \cdot n) = d_M(m, n).$$

If d_M is invariant and if the orbits are closed sets (if the orbits are not closed sets, it is possible to have $d_Q([a],[b]) = 0$ even if $[a] \neq [b]$, in this case we call d_Q a pseudo-distance. Nevertheless, this has no consequence in this paper if d_Q is only a pseudo-distance), then

$$d_Q([x],[y]) = \inf_{g \in G} d_M(x, g \cdot y),$$

is well defined, and d_Q is a distance in the quotient space. The quotient distance $d_Q([x],[y])$ is the distance between x and y' where y' is the registration of y with respect to x. We say in this case that y' is in optimal position with respect to x.

One particular distance in the ambient space M, which we use in all this article, is the distance given by the norm of the Hilbert space: $d_M(a,b) = \|a - b\|$. Moreover we say that G acts isometrically on M, if $x \mapsto g \cdot x$ is a linear map which leaves the norm unchanged. In this case d_M the distance given by the norm of the Hilbert space is invariant under the group action. The quotient (pseudo)-distance is, in this case (see fig. 2), $d_Q([a],[b]) = \inf_{g \in G} \|a - g \cdot b\|$.

Remark 1.1. *When G acts isometrically on M a Hilbert space, by expansion of the squared norm we have:*

$$d_Q([a],[b])^2 = \|a\|^2 - 2\sup_{g \in G} \langle a, g \cdot b \rangle + \|b\|^2$$

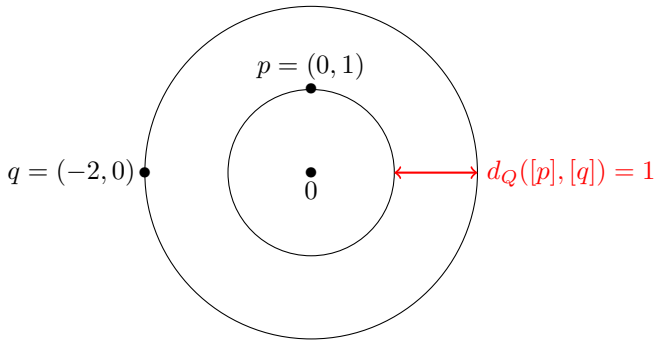


Figure 2: Due to the invariant action, the orbits are parallel. Here the orbits are circles centred at 0. This is the case when the group G is the group of rotations.

Thus, even if the quotient space is not a linear space, we have a "polarization identity" in the quotient space:

$$\sup_{g\in G} \langle a, g\cdot b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - d_Q^2([a],[b])\right) = \frac{1}{2}\left(d_Q^2([a],[0]) + d_Q^2([b],[0]) - d_Q^2([a],[b])\right) \tag{2}$$

When the distance given by the norm is invariant under the group action, we define the variance of the random orbit $[Y]$ as the expectation of the (pseudo)-distance between the random orbit $[Y]$ and the orbit of a point x in M:

$$F(x) = \mathbb{E}(d_Q^2([x],[Y])) = \mathbb{E}(\inf_{g\in G}\|g\cdot x - Y\|^2) = \mathbb{E}(\inf_{g\in G}\|x - g\cdot Y\|^2).$$

Note that $F(x)$ is well defined for all $x \in M$ because $\mathbb{E}(\|Y\|^2)$ is finite. Moreover, since $F(g\cdot x) = F(x)$, for all $x \in M$ and $g \in G$, the variance F is well defined in the quotient space: $[x] \mapsto F(x)$ does have a sense.

Moreover, in presence of a sample of the observable variable Y noted $Y_1, \ldots, Y_n$, one can define the empirical variance of a point x in M :

$$F_n(x) = \sum_{k=1}^{n}(\inf_{g\in G}\|g\cdot x - Y_i\|^2) = \sum_{k=1}^{n}(\inf_{g\in G}\|x - g\cdot Y_i\|^2).$$

Definition 1.2. *Template estimation is performed by minimizing F_n :*

$$\widehat{t_0}_n = \underset{x\in M}{argmin}\ F_n,$$

In order to study this estimation method, one can look the limit of this estimator when the number of data n tends to $+\infty$, in this case, the estimation becomes:

$$\widehat{t_0}_\infty = \underset{x\in M}{\text{argmin}}\ F$$

If $m_\star \in H$ minimizes F, then $[m_\star]$ is called a Fréchet mean of $[Y]$.

Definition 1.3. *We say that the estimation is consistent if t_0 minimizes F. Moreover the consistency bias, noted CB, is the (pseudo)-distance between the orbit of the template $[t_0]$ and $[m_\star]$: $CB = d_Q([t_0], [m_\star])$. If such a $m_\star$ does not exist, then the consistency bias is infinite.*

Note that, if the action is not isometric and is not either invariant, a priori d_Q is no longer a (pseudo)-distance in the quotient space (this point is discussed in Section 3). However one can still define F and wonder if the minimization of F is a consistent estimator of t_0. In this case, we call F a pre-variance.

1.4 Questions and Contributions

This setting leads us to wonder about few things listed below:

Questions:

- Is t_0 a minimum of the variance or the pre-variance?
- What is the behavior of the consistency bias with respect to the noise level?
- How to perform such a minimization of the variance? Indeed, in practice we have only a sample and not the whole distribution.

Contribution: In the case of an isometric action, we provide a Taylor expansion of the consistency bias when the noise level σ tends to infinity. As we do not have the whole distribution, we minimize the empirical variance given a sample. An element which minimizes this empirical variance is called an empirical Fréchet mean. We already know that the empirical Fréchet mean converges to the Fréchet mean when the sample size tends to infinity [Zie77]. Therefore our problem is reduced to finding an empirical Fréchet mean with a finite but sufficiently large sample. One algorithm called the "max-max" algorithm [AAT07] aims to compute such an empirical Fréchet mean. We establish some properties of the convergence of this algorithm. In particular, when the group is finite, the algorithm converges in a finite number of steps to an empirical Karcher mean (a local minimum of the empirical variance given a sample). This helps us to illustrate the inconsistency in this very simple framework.

We would like to insist on this point: the noise is created in the ambient space with our generative model and the computation of the Fréchet mean is done in the quotient space, this interaction induces an inconsistency. On the opposite, if one models the noise directly in the quotient space and compute the Fréchet mean in the quotient space, we have no reason to suspect any inconsistency.

Moreover it is also possible to define and use isometric actions on curves [HCG$^+$13, KSW11] or on surfaces [KKD$^+$11] where our work can be directly applied. The previous works related to the inconsistency of template estimation [BC11, MHP16, DATP17] focused on isometric action, which is a restriction to real applications. That is why we provide, in Section 3, some insights of the non invariant case: the inconsistency also appears as soon as the noise level is large enough.

This article is organized as follows: Section 2 is dedicated for isometric action. More precisely, in Section 2.2, we study the presence of the inconsistency and we establish the asymptotic behavior when the noise parameter σ tends to ∞. In Section 2.4 we detail the max-max algorithm and its properties. In Section 2.5 we illustrate the inconsistency with synthetic data. Finally in Section 3, we prove the inconsistency for more general group action, when the noise level is large enough. We do it in two settings, the first one is that the group contains a subgroup acting isometrically on M, the second one is that the group acts linearly on the space M.

2 Inconsistency of Template Estimation with an Isometric Action

2.1 Congruent Section and Computation of Fréchet Mean in Quotient Space

Given points m and y, there is a priori no closed formed expression in order to compute the quotient distance $\inf_{g\in G}\|g\cdot m-y\|$. Therefore computing and minimizing the variance in the quotient does not seem straightforward. There is one case where it may be possible: the existence of a congruent section. We say that $s : Q \to M$ is a section if $\pi\circ s = Id$, where $\pi : M \to Q$ is the canonical projection into the quotient space. Moreover we say that the section s is congruent if:

$$\forall o, o' \in Q \quad \|s(o) - s(o')\| = d_Q(o, o').$$

Then $\mathcal{S} = s(Q)$ the image of the quotient by the section is a part of M which has an interesting property:

$$\forall p, q \in \mathcal{S}, \|p - q\| = d_Q([p], [q]).$$

In other words, the section gives us a part of M containing a point of each orbit such that all points in $\mathcal{S}$ are already registered. Moreover, if s is a section, $s' : [m] \mapsto g \cdot s([m])$ is also a section, without loss of generality we can assume that $t_0 = s([t_0])$.

In this case, the variance is equal to:

$$F(m) = \mathbb{E}(\|s([m]) - s([Y])\|^2),$$

where we recognize the variance of the random variable $s([Y])$. As we know that the element which minimizes the variance in a linear space is given by the expected value, we have that:

$$F(m) \geq F(\mathbb{E}(s([Y]))).$$

Moreover this inequality is strict if and only if m and $\mathbb{E}(s([Y]))$ are not in the same orbit.

Therefore, we have a method in order to know if the estimation is consistent or not: computing $\mathbb{E}(s([Y]))$ and verifying if t_0 and $\mathbb{E}(s([Y]))$ are in

the same orbit, and the consistency bias is given by $d_Q([t_0], [\mathbb{E}(s([Y]))])$. Moreover if we take $m \in \mathcal{S}$, we have $F(m) = \mathbb{E}(\|m - s([Y])\|^2)$ and it is now straightforward that $F|_{\mathcal{S}}$ the restriction of F to $\mathcal{S}$ is differentiable on $\mathcal{S}$ (We say that $F|_{\mathcal{S}}$ is differentiable on $\mathcal{S}$, even if $\mathcal{S}$ is not open, because $m \mapsto \mathbb{E}(\|m - s([Y])\|^2)$ is defined and differentiable on M, and is equal to $F|\mathcal{S})$, and that $\nabla F|_{\mathcal{S}}(m) = m - \mathbb{E}(s([Y])$ in particular $\|\nabla F|_{\mathcal{S}}(t_0)\| = \|t_0 - \mathbb{E}(s([Y]))\|$ gives us the value of the bias.

Example 2.1. *The action of rotations: $G = SO(n)$ acts isometrically on $M = \mathbb{R}^n$. We notice that the quotient distance is $d_Q([x], [y]) = |\|x\| - \|y\||$. We can check that $s([x]) = \|x\|v$ is a section for v an unitary vector. Therefore the computation of the bias is given by $d_Q([t_0], [\mathbb{E}(s([Y])]) = |\mathbb{E}(\|Y\|) - \|t_0\|)|$.*

Unfortunately, the congruent section generally does not exist. Let us give an example:

Example 2.2. *Taking $N \in \mathbb{N}$ with $N \geq 3$, we consider the action of $G = \mathbb{Z}/N\mathbb{Z}$ on $M = \mathbb{R}^N$ by time translation, for $\bar{k} \in \mathbb{Z}/N\mathbb{Z}$, and $(x_1, x_2, \ldots, x_N)$:*

$$\bar{k} \cdot (x_1, x_2, \ldots, x_N) = (x_{1+k}, x_{2+k}, \ldots, x_{N+k}),$$

where indexes are taken modulo N. If we take $p_1 = (0, 5, 0, \ldots, 0)$, $p_2 = (0, 3, 2, 0, \ldots, 0)$, $p_3 = (2, 3, 0, \ldots, 0)$. By hand we can check that there is no $x \in [p_1]$, $y \in [p_2]$ and $z \in [p_3]$ such that $\|x - y\| = d_Q([p_1], [p_2])$, $\|x - z\| = d_Q([p_1], [p_3])$, and $\|y - z\| = d_Q([p_2], [p_3])$. Thus, a congruent section in $Q = M/G$ does not exists.

We can generalize this simple example by taking a non finite group:

Example 2.3. *Let us take $M = L^2(\mathbb{R}/\mathbb{Z})$ the set of 1-periodic functions such that $\int_0^1 f^2(t)dt < +\infty$. $G = \mathbb{R}/\mathbb{Z}$ acts on $L^2(\mathbb{R}/\mathbb{Z})$ by time translation defined by:*

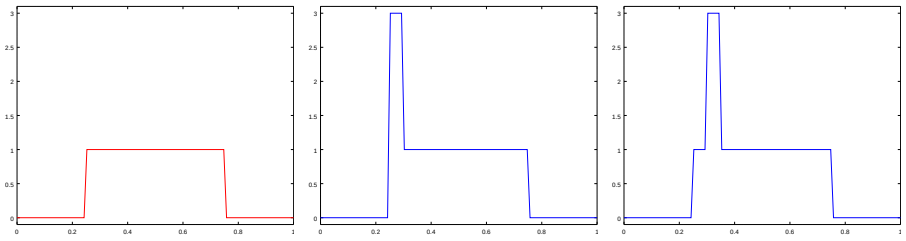
$$\tau \in \mathbb{R}/\mathbb{Z},\ f \in L^2(\mathbb{R}/\mathbb{Z}) \mapsto f_\tau \ with f(x) = f(x + \tau).$$

Then a section in $Q = M/G$ does not exists.

Proof. Let us take $f_1 = \mathbb{1}_{[\frac{1}{4}, \frac{3}{4}]}$, $f_2 = f_1 + 2\mathbb{1}_{[\frac{1}{4}, \frac{1}{4}+\eta]}$ and $f_2 = f_1 + 2\mathbb{1}_{[\frac{1}{4}+\eta, \frac{1}{4}+2\eta]}$ for some $\eta \in (0, \frac{1}{4})$ (see fig. 3). Let us suppose that a section s exists, then without loss of generality we can assume that $s([f_1]) = f_1$, then we should have $\|f_1 - s([f_2])\| = \|s([f_1]) - s([f_2])\| = d_Q([f_1], [f_2])$ in other words, $s([f_2])$ should be registered with respect to f_1. For $\tau \in \mathbb{R}/\mathbb{Z}$ we can verify that $\|f_1 - \tau \cdot f_2\| \geq \|f_1 - f_2\|$ and that this inequality is strict as soon as $\tau \neq 0$. Then f_2 is the only element of $[f_2]$ registered with f_1 then $s([f_2]) = f_2$. Likewise for $s([f_3]) = f_3$, then we should have:

$$d_Q([f_2], [f_3]) = \|f_2 - f_3\|,$$

However it is easy to verify that $d_Q^2([f_2], [f_3]) \leq \|\eta \cdot f_2 - f_3\|^2 = 2\eta < 8\eta = \|f_2 - f_3\|^2 = d_Q([f_2], [f_3])$. This is a contradiction. Therefore, a congruent section does not exist.



(a) f_1 (b) f_2 (c) f_3

Figure 3: Representation of the three functions f_1, f_2 and f_3 with $h = 0.05$. the functions f_2 and f_3 are registered with respect to f_1. However f_2 and f_3 are not registered with each other, since it is more profitable to shift f_2 in order to align the highest parts of f_2 and f_3.

When the congruent section exists, then the quotient can be included in a part $\mathcal{S}$ of the ambient space M and the metric d_M and d_Q are corresponding. The existence of a congruent section indicates us that the quotient space is not so complicated. Indeed when there is an existence of a congruent section, the quotient space is embedded in the ambient space with respect to the distances in the quotient space and in the ambient space. In that case computations are easier, projecting data on this part S and taking the mean. Then when such a congruent section does not exist, computing the Fréchet mean in quotient space is not so obvious. However, we can established proofs of inconsistency which are less tight. In this article we prove that the method is inconsistent when the noise is large.

2.2 Inconsistency and Quantification of the Consistency Bias

We start with Theorem 2.4 which gives us an asymptotic behavior of the consistency bias when the noise level σ tends to infinity. One key notion in Theorem 2.4 is the concept of fixed point under the action G: a point $x \in M$ is a fixed point if for all $g \in G$, $g \cdot x = x$. We require that the support of the noise ϵ is not included in the set of fixed points. However, this condition is almost always fulfilled. For instance in $\mathbb{R}^n$ the set of fixed points under a linear group action is a null set for the Lebesgue measure (unless the action is trivial: $g \cdot x = x$ for all $g \in G$ but this situation is irrelevant).

Theorem 2.4. *Let us suppose that the support of the noise ϵ is not included in the set of fixed points under the group action. Let Y be the observable variable defined in Equation (1). If the Fréchet mean of $[Y]$ exists, then we have the following lower and upper bounds of the consistency bias noted CB:*

$$\sigma K - 2\|t_0\| \leq CB \leq \sigma K + 2\|t_0\|, \tag{3}$$

where $K = \sup_{\|v\|=1} \mathbb{E}\left(\sup_{g\in G} \langle v, g\cdot\epsilon\rangle\right) \in (0,1]$*, K is a constant which depends only of the standardized noise and of the group action. The consistency bias has the following asymptotic behavior when the noise level σ tends to infinity:*

$$CB = \sigma K + o(\sigma) \text{ as } \sigma \to +\infty. \tag{4}$$

In the following we note by S the unit sphere of M. For $v \in S$, we call $\theta(v) = \mathbb{E}\left(\sup_{g\in G} \langle v, g\cdot\epsilon\rangle\right)$, so that $K = \sup_{v\in S} \theta(v)$. The sketch of the proof is the following:

- $K > 0$ because the support of ϵ is not included in the set of fixed points under the action of G.
- $K \leq 1$ is the consequence of the Cauchy-Schwarz inequality.
- The proof of Inequalities (3) is based on the triangular inequalities:

  $$\|m_\star\| - \|t_0\| \leq CB = \inf_{g\in G} \|t_0 - g\cdot m_\star\| \leq \|t_0\| + \|m_\star\|,$$

  where $m_\star$ minimizes F: having a piece of information about the norm of $m_\star$ is enough to deduce a piece of information about the consistency bias.
- The asymptotic Taylor expansion of the consistency bias (4) is the direct consequence of inequalities (3).

Proof of Theorem 2.4. We note S the unit sphere in M. In order to prove that $K > 0$, we take x in the support of ϵ such that x is not a fixed point under the action of G. It exists $g_0 \in G$ such that $g_0 \cdot x \neq x$. We note $v_0 = \frac{g_0\cdot x}{\|x\|} \in S$, we have $\langle v_0, g_0\cdot x\rangle = \|x\| > \langle v_0, x\rangle$ and by continuity of the dot product it exists $r > 0$ such that: $\forall y \in B(x,r) \quad \langle v_0, g_0\cdot y\rangle > \langle v_0, y\rangle$ as x is in the support of ϵ we have $\mathbb{P}(\epsilon \in B(x,r)) > 0$, it follows:

$$\mathbb{P}\left(\sup_{g\in G} \langle v_0, g\cdot\epsilon\rangle > \langle v_0, \epsilon\rangle\right) > 0. \tag{5}$$

Thanks to Inequality (5) and the fact that $\sup_{g\in G} \langle v_0, g\cdot\epsilon\rangle \geq \langle v_0, \epsilon\rangle$ we have:

$$\theta(v_0) = \mathbb{E}\left(\sup_{g\in G} \langle v_0, g\cdot\epsilon\rangle\right) > \mathbb{E}(\langle v_0, \epsilon\rangle) = \langle v_0, \mathbb{E}(\epsilon)\rangle = \langle v_0, 0\rangle = 0.$$

Then we get $K \geq \theta(v_0) > 0$. Moreover, if we use the Cauchy-Schwarz inequality:

$$K \leq \sup_{v\in S} \mathbb{E}(\|v\| \times \|\epsilon\|) \leq \mathbb{E}(\|\epsilon\|^2)^{\frac{1}{2}} = 1.$$

In order to prove Inequalities (3), we use the "polar" coordinates of a point in M (see fig. 4), every point in M can be represented by (r, v) where $r \geq 0$ is the radius, and v belong to S the unit sphere in M, v represents the "angle". We compute $F(m)$ as a function of (r, v). In a first step, we minimize this expression as a function of r, in a second step we minimize this expression as a function of v. This makes appear the constant K.

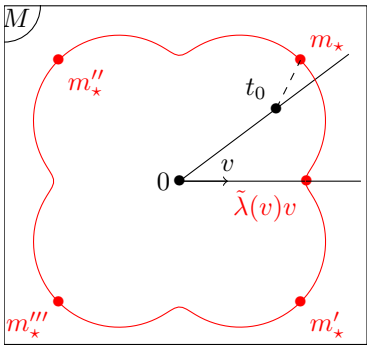


Figure 4: We minimize the variance on each half-line $\mathbb{R}^+ v$ where $\|v\| = 1$. The element which minimizes the variance on such a half-line is $\tilde{\lambda}(v)v$, where $\tilde{\lambda}(v) \geq 0$. We get a surface in M by $S \in v \mapsto \tilde{\lambda}(v)v$ (which is a curve in this figure since we draw it in dimension 2). The Proof of Theorem 2.4 states that if $[m_\star]$ is a Fréchet mean then $m_\star$ is an extreme point of this surface. On this picture there are four extreme points which are in the same orbit: we took here the simple example of the group of rotations of 0, 90, 180 and 270 degrees.

As we said, let us take $r \geq 0$ and $v \in S$, we expand the variance at the point rv:

$$F(rv) = \mathbb{E}\left(\inf_{g\in G} \|rv - g \cdot Y\|^2\right) = r^2 - 2r\mathbb{E}\left(\sup_{g\in G} \langle v, g \cdot Y\rangle\right) + \mathbb{E}(\|Y\|^2). \tag{6}$$

Indeed $\|g \cdot Y\| = \|Y\|$ thanks to the isometric action. We note $x^+ = \max(x, 0)$ the positive part of x. Moreover we define the two following functions:

$$\lambda(v) = \mathbb{E}(\sup_{g\in G} \langle v, g \cdot Y\rangle) = \mathbb{E}(\sup_{g\in G} \langle g \cdot Y, v\rangle) \text{ and } \tilde{\lambda}(v) = \lambda(v)^+ \text{ for } v \in S,$$

since that $f : x \in \mathbb{R}^+ \mapsto x^2 - 2bx + c$ reaches its minimum at the point $r = b^+$ and $f(b^+) = c - (b^+)^2$, the $r_\star \geq 0$ which minimizes (6) is $\tilde{\lambda}(v)$ and the minimum value of the variance restricted to the half line $\mathbb{R}^+ v$ is:

$$F(\tilde{\lambda}(v)v) = \mathbb{E}(\|Y\|^2) - \tilde{\lambda}(v)^2.$$

To find $[m_\star]$ the Fréchet mean of $[Y]$, we need to maximize $\tilde{\lambda}(v)^2$ with respect to $v \in S$:

$$m_\star = \lambda(v_\star)v_\star \text{ with } v_\star \in \operatorname*{argmax}_{v\in S} \lambda(v).$$

Note that we remove the positive part and the square because $\operatorname{argmax}\lambda = \operatorname{argmax}(\lambda^+)^2$ indeed λ takes a non negative value. In order to prove it let us remark that:

$$\lambda(v) \geq \mathbb{E}(\langle v, \Phi \cdot t_0 + \epsilon\rangle) = \langle v, \mathbb{E}(\Phi \cdot t_0)\rangle + 0,$$

then there is two cases: if $\mathbb{E}(\Phi \cdot t_0) = 0$ then for any $v \in S$ we have $\lambda(v) \geq 0$, if $w = \mathbb{E}(\Phi \cdot t_0) \neq 0$ then we take $v = \frac{w}{\|w\|} \in S$, and we get $\lambda(v) \geq \left\langle \frac{w}{\|w\|}, w \right\rangle = \|w\| \geq 0$.

As we said in the sketch of the proof we are interested in getting information about the norm of $\|m_\star\|$:

$$\|m_\star\| = \lambda(v_\star) = \sup_{v \in S} \lambda.$$

Let $v \in S$, we have: $-\|t_0\| \leq \langle v, g\Phi \cdot t_0 \rangle \leq \|t_0\|$ because the action is isometric. Now we decompose $Y = \Phi \cdot t_0 + \sigma\epsilon$ and we get:

$$\lambda(v) = \mathbb{E}\left(\sup_{g \in G} \langle v, g \cdot Y \rangle\right) = \mathbb{E}\left(\sup_{g \in G} \left(\langle v, g \cdot \sigma\epsilon \rangle + \langle v, g\Phi \cdot t_0 \rangle\right)\right) \tag{7}$$

$$\lambda(v) \leq \mathbb{E}\left(\sup_{g \in G} \left(\langle v, g \cdot \sigma\epsilon \rangle + \|t_0\|\right)\right) = \sigma\mathbb{E}\left(\sup_{g \in G} \langle v, g \cdot \epsilon \rangle\right) + \|t_0\| \tag{8}$$

$$\lambda(v) \geq \mathbb{E}\left(\sup_{g \in G} \left(\langle v, g \cdot \sigma\epsilon \rangle\right) - \|t_0\|\right) = \sigma\mathbb{E}\left(\sup_{g \in G} \langle v, g \cdot \epsilon \rangle\right) - \|t_0\|. \tag{9}$$

By taking the largest value in these inequalities with respect to $v \in S$, we get by definition of K:

$$-\|t_0\| + \sigma K \leq \|m_\star\| = \sup_{v \in S} \lambda(v) \leq \|t_0\| + \sigma K. \tag{10}$$

Moreover we recall the triangular inequalities:

$$\|m_\star\| - \|t_0\| \leq CB = \inf_{g \in G} \|t_0 - g \cdot m_\star\| \leq \|t_0\| + \|m_\star\|, \tag{11}$$

Thanks to (10) and to (11), Inequalities (3) are proved.

2.3 Remarks about Theorem 2.4 and Its Proof

We can ensure the presence of inconsistency as soon as the signal to noise ratio satisfies $\frac{\|t_0\|}{\sigma} < \frac{K}{2}$. Moreover, if the signal to noise ratio verifies $\frac{\|t_0\|}{\sigma} < \frac{K}{3}$ then the consistency bias is not smaller than $\|t_0\|$ i.e.,: $CB \geq \|t_0\|$. In other words, the Fréchet mean in quotient space is too far from the template: the template estimation with the Fréchet mean in quotient space is useless in this case. In [DATP17] we also gave lower and upper bounds as a function of σ but these bounds were less informative than bounds given by Theorem 2.4. These bounds did not give the asymptotic behaviour of the consistency bias. Moreover, in [DATP17] the lower bound goes to zero when the template becomes closed to fixed points. This may suggest that the consistency bias was small for this kind of template. We prove here that it is not the case.

Note that Theorem 2.4 is not a contradiction with [KSW11] where the authors proved the consistency of template estimation with the Fréchet mean in quotient space for all $\sigma > 0$. Indeed their noise was included in the set of constant functions which are the fixed points under their group action.

The constant K appearing in the asymptotic behaviour of the consistency bias (4) is a constant of interest. We can give several (but similar) interpretations of K:

- It follows from Equation (3) that K is the consistency bias with a null template $t_0 = 0$ and a standardized noise ($\sigma = 1$).
- From the proof of Theorem 2.4 we know that $0 < K \leq \mathbb{E}(\|\epsilon\|) \leq 1$. On the one hand, if G is the group of rotations then $K = \mathbb{E}(\|\epsilon\|)$, because for all v s.t. $\|v\| = 1$, $\sup_{g\in G} \langle v, g\epsilon\rangle = \|\epsilon\|$, by aligning v and ϵ. On the other hand if G acts trivially (which means that $g \cdot x = x$ for all $g \in G$, $x \in M$) then $K = 0$. The general case for K is between two extreme cases: the group where the orbits are minimal (one point) and the group for which the orbits are maximal (the whole sphere). We can state that the more the group action has the ability to align the elements, the larger the constant K is and the larger the consistency bias is.
- The squared quotient distance between two points is:

$$d_Q([a],[b])^2 = \|a\|^2 - 2\sup_{g\in G} \langle a, g \cdot b\rangle + \|b\|^2,$$

  thus the larger $\sup_{g\in G} \langle a, g \cdot b\rangle$, the smaller $d_Q([a],[b])$. $K = 1 - \frac{1}{2}\inf_{\|v\|=1} \mathbb{E}(d_Q^2([v],[\epsilon]))$, encodes the level of contraction of the quotient distance (or folding). The larger K is, the more contracted the quotient space is.

One disadvantage of Theorem 2.4 is that it ensures the presence of inconsistency for σ large enough but it says nothing when σ is small, in this case one can refer to [MHP16] or [DATP17].

We can remark that this Theorem can be used as an alternating proof the following Theorem (which was already proved in [DATP17]), proving and quantifying inconsistency when the template is a fixed point:

Corollary 2.5. *Let G acting isometrically on M an Hilbert space. Let t_0 be a fixed point, and ϵ a standardized noise which support is not included in the set of fixed points. Then estimating the template with the Fréchet mean is inconsistent. Moreover if the Fréchet mean in quotient space exists then the consistency bias is equal to:*

$$CB = \sigma K.$$

Indeed for $t_0 = 0$ which is a particular fixed point we have $CB = \sigma K$ thanks to Theorem 2.4. If t_0 is a fixed point non necessarily equal to 0, we can define $Y' = Y - t_0 = 0 + \sigma\epsilon$, in this random variable 0 is the template we can apply the formula $CB = \sigma K$ to the random variable Y', which concludes.

In the proof of Theorem 2.4, we have seen that the minimum of the variance restricted to the half-line $\mathbb{R}^+ v$ for $v \in S$, was

$$\mathbb{E}(\|Y\|^2) - \left(\left(\mathbb{E}\left(\inf_{g\in G} \langle v, g \cdot Y\rangle\right)\right)^+\right)^2.$$

therefore $\tilde{\lambda}(v) = \left(\mathbb{E}\left(\inf_{g\in G} \langle v, g \cdot Y\rangle\right)\right)^+$ is a registration score: $\tilde{\lambda}(v)$ tells you how much it is a good idea to search the Fréchet mean of $[Y]$ in the

direction pointed by v: the more $\tilde{\lambda}(v)$ is large, the more v is a good choice. On the contrary when this value is equal to zero, it is useless to search the Fréchet mean in this direction.

Likewise, for $v \in S$, $\theta(v) = \mathbb{E}(\sup_{g\in G} \langle g \cdot v, \epsilon\rangle)$ is a registration score with respect to the noise, the larger $\theta(v)$, the more the unit vector v looks like to the noise ϵ after registration.

If $[m_\star]$ is a Fréchet mean of $[Y]$ we have seen that its norm verifies:

$$\|m_\star\| = \sup_{\|v\|=1} \mathbb{E}(\sup_{g\in G} \langle v, g \cdot Y\rangle).$$

Then if there is two different Fréchet means of $[Y]$ noted $[m_\star]$ and $[n_\star]$, we can deduce that $\|m_\star\| = \|n_\star\|$. Even if there is no uniqueness of the Fréchet mean in the quotient space, we can state that the representants of the different Fréchet means have all the same norm.

Remark 2.6. *We can also wonder if the converse of Theorem 2.4 is true: if ϵ is a non biased noise always included in the set of fixed point, is $[t_0]$ a Fréchet mean of $[\Phi \cdot t_0 + \sigma\epsilon]$? A simple computation show that t_0 is a minimum of the variance:*

$$\begin{aligned} F(m) &= \mathbb{E}\left(\inf_{g\in G} \|m - g \cdot (\Phi t_0 + \sigma\epsilon)\|^2\right) \\ &= \|m\|^2 + \mathbb{E}(\|\Phi t_0 + \sigma\epsilon\|^2) - 2\mathbb{E}(\sup_g \langle m, g\Phi t_0\rangle + \langle m, g\sigma\epsilon\rangle) \\ &= \|m\|^2 + \mathbb{E}(\|\Phi t_0 + \sigma\epsilon\|^2) - 2\mathbb{E}\left(\sup_{g\in G} \langle m, g \cdot t_0\rangle\right) - 2\langle m, \mathbb{E}(\sigma\epsilon)\rangle \\ &= \|m\|^2 + \mathbb{E}(\|\Phi t_0 + \sigma\epsilon\|^2) - 2\mathbb{E}\left(\sup_{g\in G} \langle m, g \cdot t_0\rangle\right) \qquad (12) \end{aligned}$$

We see that the element m which minimizes (12) *does not depend of σ, in particular we can assume $\sigma = 0$, and wonder which elements minimizes $F(m) = \mathbb{E}(\inf_{g\in G} \|m - g\Phi \cdot t_0\|^2)$, it becomes clear that only the points in the orbit of t_0 can minimize this variance. Then when ϵ is included in the set of fixed points, the estimation is always consistent for all σ. This is an alternative proof of the Theorem of consistency done by Kurtek et al. [KSW11].*

In the proof of Theorem 2.4, we have seen that the direction of the Fréchet mean of $[Y]$ is given by the supremum of this quantity (7):

$$\mathbb{E}\left(\sup_{g\in G} \langle v, g \cdot \sigma\epsilon\rangle + \langle v, g\Phi \cdot t_0\rangle\right).$$

This Equation is a good illustration of the difficulty to compute the Fréchet mean in quotient space. Indeed, we have on one side the contribution of the noise $\langle v, g \cdot \sigma\epsilon\rangle$ and on the other side the contribution of the template $\langle v, g\Phi \cdot t_0\rangle$, and we take the supremum of the sum of these two contributions over $g \in G$. Unfortunately the supremum of the sum of two terms is not equal to the sum of the supremum of each of these terms. Hence, it is difficult to separate these two contributions. However, we can

intuit that when the noise is large, $\langle v, g\sigma\epsilon\rangle$ prevails over $\langle v, g\Phi \cdot t_0\rangle$, and the use of the Cauchy-Schwarz inequality in Equations (8) and (9) proves it rigorously. We can conclude that, when the noise is large, the direction of the Fréchet mean in the quotient space depends more on the noise than on the template.

2.4 Template Estimation with the Max-Max Algorithm

2.4.1 Max-Max Algorithm Converges to a Local Minima of the Empirical Variance

Section 2.2 can be understood as follows: if we want to estimate the template by minimizing the Fréchet mean with quotient space, then there is a bias. This supposes that we are able to compute such a Fréchet mean. In practice, we cannot minimize the exact variance in quotient space, because we have only a finite sample and not the whole distribution. In this section we study the estimation of the empirical Fréchet mean with the max-max algorithm. We assume that the group is finite. In this case, the registration can always be found by an exhaustive search. Hence, the numeric experiments which we conduct in Section 2.5 lead to an empirical Karcher mean in a finite number of steps. In a compact group acting continuously, the registration also exists but is not necessarily computable without approximation.

If we have a sample: $Y_1, \ldots, Y_I$ of independent and identically distributed copies of Y, then we define the empirical variance in the quotient space:

$$M \ni x \mapsto F_I(x) = \frac{1}{I}\sum_{i=1}^{I} d_Q^2([x],[Y_i]) = \frac{1}{I}\sum_{i=1}^{I} \min_{g_i \in G} \|x - g_i \cdot Y_i\|^2 = \frac{1}{I}\sum_{i=1}^{I} \min_{g_i \in G} \|g_i \cdot x - Y_i\|^2. \tag{13}$$

The empirical variance is an approximation of the variance. Indeed thanks to the law of large number we have $\lim_{I\to\infty} F_I(x) = F(x)$ for all $x \in M$. One element which minimizes globally (respectively locally) F_I is called an empirical Fréchet mean (respectively an empirical Karcher mean). For $x \in M$ and $\underline{g} \in G^I$: $\underline{g} = (g_1, \ldots, g_I)$ where $g_i \in G$ for all $i = 1..I$ we define J an auxiliary function by:

$$J(x, \underline{g}) = \frac{1}{I}\sum_{i=1}^{I} \|x - g_i \cdot Y_i\|^2 = \frac{1}{I}\sum_{i=1}^{I} \|g_i^{-1} \cdot x - Y_i\|^2.$$

The max-max algorithm 1 iteratively minimizes the function J in the variable $x \in M$ and in the variable $\underline{g} \in G^I$ (see fig. 5):

First, we note that this algorithm is sensitive to the the starting point. However we remark that $m_1 = \frac{1}{I}\sum_{i=1}^{I} g_i \cdot Y_i$ for some $g_i \in G$, thus without loss of generality, we can start from $m_1 = \frac{1}{I}\sum_{i=1}^{I} g_i \cdot Y_i$ for some $g_i \in G$. The empirical variance does not increase at each step of the algorithm since:

$$F_I(m_n) = J(m_n, \underline{g}^n) \geq J(m_{n+1}, \underline{g}^n) \geq J(m_{n+1}, \underline{g}^{n+1}) = F_I(m_{n+1})$$

Algorithm 1 Max-Max Algorithm

Require: A starting point $m_0 \in M$, a sample $Y_1, \dots, Y_I$.

$n = 0$.

while Convergence is not reached **do**

Minimizing $\underline{g} \in G^I \mapsto J(m_n, \underline{g})$: we get g_i^n by registering Y_i with respect to m_n.

Minimizing $x \in M \mapsto J(x, \underline{g}^n)$: we get $m_{n+1} = \frac{1}{I} \sum_{i=1}^{I} g_i^n Y_i$.

$n = n + 1$.

end while

$\hat{m} = m_n$

Proposition 2.7. *As the group is finite, the convergence is reached in a finite number of steps.*

Proof of Proposition 2.7. The sequence $(F_I(m_n))_{n\in\mathbb{N}}$ is non-increasing. Moreover the sequence $(m_n)_{n\in\mathbb{N}}$ takes value in a finite set which is: $\{\frac{1}{I}\sum_{i=1}^{I} g_i \cdot Y_i, g_i \in G\}$. Therefore, the sequence $(F_I(m_n))_{n\in\mathbb{N}}$ is stationary. Let $n \in \mathbb{N}$ such that $F_I(m_n) = F_I(m_{n+1})$. Hence the empirical variance did not decrease between step n and step $n+1$ and we have:

$$F_I(m_n) = J(m_n, \underline{g}_n) = J(m_{n+1}, \underline{g}_n) = J(m_{n+1}, \underline{g}_{n+1}) = F_I(m_{n+1}),$$

as m_{n+1} is the unique element which minimizes $m \mapsto J(m, \underline{g}_n)$ we conclude that $m_{n+1} = m_n$.

This proposition gives us a shutoff parameter in the max-max algorithm: we stop the algorithm as soon as $m_n = m_{n+1}$. Let us call $\hat{m}$ the final result of the max-max algorithm. It may seem logical that $\hat{m}$ is at least a local minimum of the empirical variance. However this intuition may be wrong: let us give a toy counterexample, suppose that we observe $Y_1, \dots, Y_I$, due to the transformation of the group it is possible that $\sum_{i=1}^{n} Y_i = 0$. We can start from $m_1 = 0$ in the max-max algorithm, as Y_i and 0 are already registered, the max-max algorithm does not transform Y_i. At step two, we still have $m_2 = 0$, by induction the max-max algorithm stays at 0 even if 0 is not a Fréchet or Karcher mean of $[Y]$. Because 0 is equally distant from all the points in the orbit of Y_i, 0 is called a focal point of $[Y_i]$. The notion of focal point is important for the consistency of the Fréchet mean in manifold [BB08]. Fortunately, the situation where $\hat{m}$ is not a Karcher mean is almost always avoided due to the following statement:

Proposition 2.8. *Let $\hat{m}$ be the result of the max-max algorithm. If the registration of Y_i with respect to $\hat{m}$ is unique, in other words, if $\hat{m}$ is not a focal point of Y_i for all $i \in 1..I$ then $\hat{m}$ is a local minimum of F_I: $[\hat{m}]$ is an empirical Karcher mean of $[Y]$.*

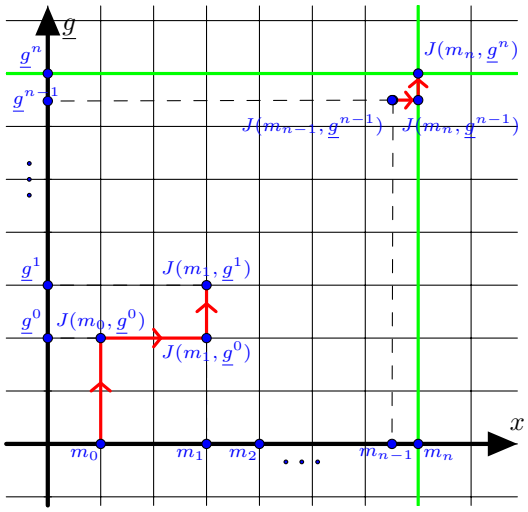


Figure 5: Iterative minimization of the function J on the two axis, the horizontal axis represents the variable in the space M, the vertical axis represents the set of all the possible registrations G^I. Once the convergence is reached, the point (m_n, g_n) is the minimum of the function J on the two axis in green. Is this point the minimum of J on its whole domain? There are two pitfalls: firstly this point could be a saddle point, it can be avoided with Proposition 2.8, secondly this point could be a local (but not global) minimum, this is discussed in Section 2.5.3

Note that, if we call z the registration of y with respect to m, then the registration is unique if and only if $\langle m, z - g \cdot z \rangle \neq 0$ for all $g \in G \setminus \{e\}$. Once the max-max algorithm has reached convergence, it suffices to test this condition for $\hat{m}$ obtained by the max-max algorithm and Y_i for all i. This condition is in fact generic and is always obtained in practice.

Proof of Proposition 2.8. We call g_i the unique element in G which register Y_i with respect to $\hat{m}$, for all $h \in G \setminus \{g_i\}$, $\|\hat{m} - g_i \cdot Y_i\| < \|\hat{m} - h_i \cdot Y_i\|$. By continuity of the norm we have for a close enough to m: $\|a - g_i \cdot Y_i\| < \|a - h_i \cdot Y_i\|$ for all $h_i \neq g_i$ (note that this argument requires a finite group). The registrations of Y_i with respect to m and to a are the same:

$$F_I(a) = \frac{1}{I} \sum_{i=1}^{I} \|a - g_i \cdot Y_i\|^2 = J(a, \underline{g}) \geq J(\hat{m}, \underline{g}) = F_I(\hat{m}),$$

because $m \mapsto J(m, \underline{g})$ has one unique local minimum $\hat{m}$.

Remark 2.9. *We remark the max-max algorithm is in fact a gradient descent. The gradient descent is a general method to find the minimum of a differentiable function. Here we are interested in the minimum of the variance F: let $m_0 \in M$ and we define by induction the gradient descent of the variance $m_{n+1} = m_n - \rho \nabla F(m_n)$, where $\rho > 0$ and F the*

variance in the quotient space. In [DATP17], the gradient of the variance in quotient space for finite group and for a regular point m was computed (m is regular as soon as $g \cdot m = m$ implies $g = e$), this leads to:

$$m_{n+1} = m_n - 2\rho \left[m_n - \mathbb{E}(g(Y, m_n) \cdot Y) \right],$$

where $g(Y, m_n)$ is the almost-surely unique element of the group which registers Y with respect to m_n. Now if we have a set of data $Y_1, \ldots, Y_n$ we can approximated the expectation which leads to the following approximated gradient descent:

$$m_{n+1} = m_n(1 - 2\rho) + \rho \frac{2}{I} \sum_{i=1}^{I} g(Y_i, m_n) \cdot Y_i,$$

now by taking $\rho = \frac{1}{2}$ we get $m_{n+1} = \frac{1}{I} \sum_{i=1}^{I} g(Y_i, m_n) \cdot Y_i$. So the approximated gradient descent with $\rho = \frac{1}{2}$ is exactly the max-max algorithm. However, the max-max algorithm for finite group, is proved to be converging in a finite number of steps which is not the case for gradient descent in general.

2.5 Simulation on Synthetic Data

In this Section, we consider data in an Euclidean space $\mathbb{R}^N$ equipped with its canonical dot product $\langle \cdot, \cdot \rangle$, and $G = \mathbb{Z}/\mathbb{N}\mathbb{Z}$ acts on $\mathbb{R}^N$ by time translation on coordinates:

$$(\bar{k} \in \mathbb{Z}/N\mathbb{Z}, (x_1, \ldots, x_N) \in \mathbb{R}^N) \mapsto (x_{1+k}, x_{2+k}, \ldots x_{N+k}),$$

where indexes are taken modulo N. This space models the discretization of functions defined on $[0, 1]$ with N points. This action is found in [AAT07] and used for neuroelectric signals in [HCG+13]. The registration between two vectors can be made by an exhaustive research but it is faster with the fast Fourier transform [CT65].

2.5.1 Max-Max Algorithm with a Step Function as Template

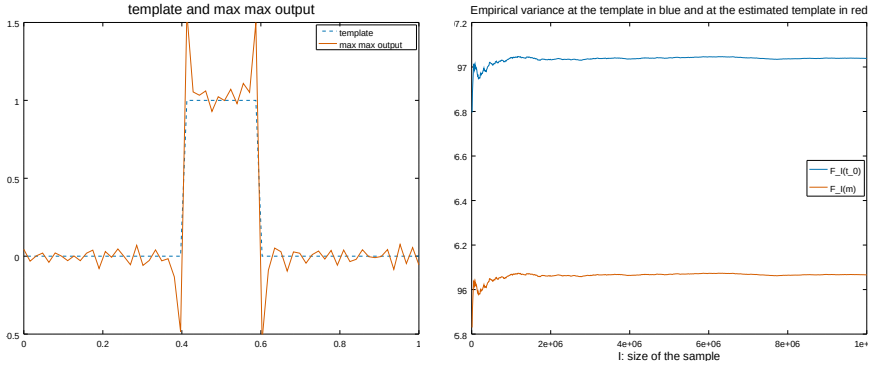
We display an example of a template and template estimation with the max-max algorithm on Figure 6a. This experiment was already conducted in [AAT07], but no explanation of the appearance of the bias was provided. We know from Section 2.4 that the max-max output is an empirical Karcher mean, and that this result can be obtained in a finite number of steps. Taking $\sigma = 10$ may seem extremely high, however the standard deviation of the noise at each point is not 10 but $\frac{\sigma}{\sqrt{N}} = 1.25$ which is reasonable.

The sample size is 10^5, the algorithm stopped after 247 steps, and $\hat{m}$ the estimated template (in red on the Figure 6a) is not a focal points of the orbits $[Y_i]$, then Proposition 2.8 applies. We call empirical bias (noted EB) the quotient distance between the true template and the point $\hat{m}$ given by the max-max result. On this experiment we have $\frac{EB}{\sigma} \simeq 0.11$. Of course, one could think that we estimate the template with an empirical bias due

to a too small sample size which induces fluctuation. To reply to this objection, we keep in memory $\hat{m}$ obtained with the max-max algorithm. If there was no inconsistency then we would have $F(t_0) \leq F(\hat{m})$. We do not know the value of the variance F at these points, but thanks to the law of large number, we know that:

$$F(t_0) = \lim_{I \to \infty} F_I(t_0) \text{ and } F(\hat{m}) = \lim_{I \to \infty} F_I(\hat{m}),$$

Given a sample, we compute $F_I(t_0)$ and $F_I(\hat{m})$ thanks to the definition of the empirical variance F_I (13). We display the result on Figure 6b, this tends to confirm that $F(t_0) > F(\hat{m})$. In other word, the variance at the template is larger that the variance at the point given by the max-max algorithm.



(a) (b)

Figure 6: Template t_0 and template estimation $\hat{m}$ on Figure 6a. Empirical variance at the template and template estimation with the max-max algorithm as a function of the size of the sample on Figure 6b. (**a**) Example of a template (a step function) and the estimated template $\hat{m}$ with a sample size 10^5 in $\mathbb{R}^{64}$, ϵ is Gaussian noise and $\sigma = 10$. At the discontinuity points of the template, we observe a Gibbs-like phenomena; (**b**) Variation of $F_I(t_0)$ (in blue) and of $F_I(\hat{m})$ (in red) as a function of I the size of the sample. Since convergence is already reached, $F(\hat{m})$, which is the limit of red curve, is below $F(t_0)$: $F(t_0)$ is the limit of the blue curve. Due to the inconsistency, $\hat{m}$ is an example of point such that $F(\hat{m}) < F(t_0)$.

2.5.2 Max-Max Algorithm with a Continuous Template

Figure 6a shows that the main source of the inconsistency was the discontinuity of the template. One may think that a continuous template would lead to a better behaviour. However, it is not the case as presented on Figure 7. Even with a large number of observations created from a continuous template we do not observe a convergence to the template. In the example of Figure 7, the empirical bias satisfies $\frac{EB}{\sigma} = 0.23$. In green

we also display the mean of data knowing transformations, this produces a much better result, since that in this case we have $\frac{EB}{\sigma} = 0.04$.

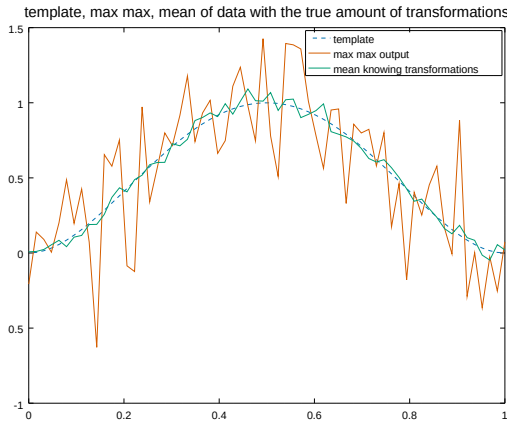


Figure 7: Example of an other template (here a discretization of a continuous function) and template estimation with a sample size 10^3 in $\mathbb{R}^{64}$, ϵ is Gaussian noise and $\sigma = 10$. Even with a continuous function the inconsistency appears.

2.5.3 Does the Max-Max Algorithm Give Us a Global Minimum or Only a Local Minimum of the Variance?

Proposition 2.8 tells us that the output of the max-max algorithm is a Karcher mean of the variance, but we do not know whether it is Fréchet mean of the variance. In other words, is the output a global minimum of the variance? In fact, F_I has a lot of local minima which are not global. To illustrate this, we may use the max-max algorithm with different starting points and we observe different outputs (which are all local minima thanks to Proposition 2.8) with different empirical variance on Table 1.

Points	**Template t_0**	$\hat{m}_1$	$\hat{m}_2$	$\hat{m}_3$	$\hat{m}_4$	$\hat{m}_5$
Empirical variance at these points	96.714	95.684	95.681	95.676	95.677	95.682

Table 1: Empirical variances at 5 different outputs of the max-max algorithm coming from the same sample of size 10^4, but with different starting points. We use $\sigma = 10$ and the action of time translation in $\mathbb{R}^{64}$. Conclusion: on these five points, only $\hat{m}_3$ is an eventual global minima.

3 Inconsistency in the Case of Non Invariant Distance under the Group Action

3.1 Notation and Hypothesis

In this Section, data still come from an Hilbert space M. However, we take a group of deformation G which acts in a non invariant way on M. Starting from a template t_0 we consider a random deformation in the group G namely a random variable Φ which takes value in G and ϵ an standardized noise in M independent of Φ. We suppose that our observable random variable is:

$$Y = \Phi \cdot t_0 + \sigma\epsilon \text{ with } \sigma > 0,\ \mathbb{E}(\epsilon) = 0,\ \mathbb{E}(\|\epsilon\|^2) = 1,$$

where σ is the noise level. We suppose that $\mathbb{E}(\|Y\|^2) < +\infty$, and we define the pre-variance of Y in M/G as the map defined by:

$$F(m) = \mathbb{E}\left(\inf_{g\in G} \|g \cdot m - Y\|^2\right).$$

In this part we still study the inconsistency of template estimation by minimizing F.

We present two frameworks where we can ensure the presence of inconsistency: in Section 3.3 we suppose that the group G contains a non trivial group H which acts isometrically on M. However, some groups do not satisfy this hypothesis, that is why, in Section 3.4 we do not suppose that G contains a subgroup acting isometrically but we require that G acts linearly on M. In both sections we prove inconsistency as soon as the variance σ^2 is large enough.

These hypothesis are not unacceptable as for example, deformations that are considered in computational anatomy may include rotations which form a subgroup H of the diffeomorphic deformations which acts isometrically. Concerning the second case, an important example is:

Example 3.1. *Let G be a subgroup of the group of C^∞ diffeomorphisms on $\mathbb{R}^n$ G acts linearly on $L^2(\mathbb{R}^n)$ with the map:*

$$\forall \varphi \in G \quad \forall f \in L^2(\mathbb{R}^n) \qquad \varphi \cdot f = f \circ \varphi^{-1}.$$

Note that this action is not isometric: indeed, $f \circ \varphi^{-1}$ has generally a different L^2-norm than f, because a Jacobian determinant appears in the computation of the integral.

3.2 Where Did We Need an Isometric Action Previously?

Let M be an Hilbert space, and G a group acting on M. Can we define a distance in the quotient space $Q = M/G$ defined as the set which contains all the orbits? When the action is invariant, the orbits are parallel in the sense where $d_M(m, n) = d_M(g\cdot m, g\cdot n)$ for all $m, n \in M$ and for all $g \in G$. This implies that:

$$d_Q([m], [n]) = \inf_{g\in G} \|m - g \cdot n\|,$$

is a distance on Q. However, it is not necessarily the case when the action is no longer invariant. Let us take the following example:

Example 3.2. *We call $C^\infty_{diff}(\mathbb{R}^2)$ the set of the C^∞ diffeomorphisms of $\mathbb{R}^2$. We equip $\mathbb{R}^2$ with its canonical Euclidean structure. We take $p = (-1,-1)$, $q = (1,1)$ and $r = (2,0)$ (see fig. 8),*

$$G = \left\{ f \in C^\infty_{diff}(\mathbb{R}^2) \mid f(q) = (q),\ f(p) = (p),\ \forall x \in \mathbb{R}\ f(x,0) \in \mathbb{R}r \right\}, \tag{14}$$

G acts on $\mathbb{R}^2$ by $f \cdot (x,y) = f(x,y)$. Then q and p are fixed points under this group action and the orbit of r is the horizontal line $\{(x,0), x \in \mathbb{R}\}$. On this example:

$$\inf_{g\in G} \|q - g\cdot r\| = \|q - (1,0)\| = 1 \qquad \text{however} \qquad \inf_{g\in G} \|r - g\cdot q\| = \|r - q\| = \sqrt{2},$$

then the function d_Q is not symmetric. One could think define a distance by:

$$\tilde{d}_Q([m],[n]) = \inf_{h,g\in G} \|h\cdot m - g\cdot n\|.$$

Unfortunately, in this case $\tilde{d}_Q([p],[q]) = \|p - q\| = 2\sqrt{2}$ and $\tilde{d}_Q([p],[r]) = 1 = \tilde{d}_Q([r],[q])$ then we do not have $\tilde{d}_Q([p],[q]) \leq \tilde{d}_Q([p],[r]) + \tilde{d}_Q([r],[q])$. In other words we do not have the triangular inequality.

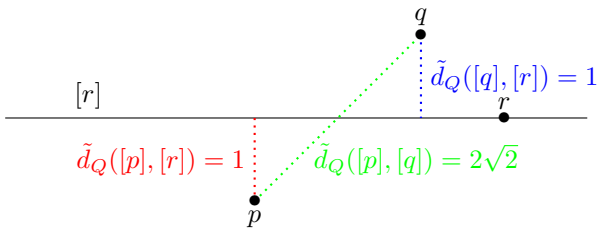


Figure 8: Example of three orbits, when $\tilde{d}_Q$ does not satisfy the inequality triangular.

Therefore when the action is no longer invariant, a priori one cannot define a distance in the quotient anymore. If Y is a random variable in M, $F(m) = \mathbb{E}(\inf_{g\in G} \|g\cdot m - Y\|^2)$ cannot be interpreted as the variance of $[Y]$.

However $\inf_{g\in G} \|g\cdot a - b\|$ is positive and is equal to zero if $a = b$, then $\inf_{g\in G} \|g\cdot a - b\|$ is a pre-distance in M. Then $\inf_{g\in G} \|g\cdot m - Y\|$ measures the discrepancy between the random point Y and the current point m. Even if the discrepancy measure is not symmetric or does not satisfy the triangular inequality, we can still define $F(x) = \mathbb{E}(\inf_{g\in G} \|g\cdot x - Y\|^2)$ and call it the pre-variance of the projection of Y into M/G, if $\mathbb{E}(\|Y\|^2) < +\infty$. The elements which minimize this function are the element whose orbit are the closest of the random point Y. Hence, we wonder if the template can be estimated by minimizing this pre-variance. Note that, once again

$F(x) = F(g \cdot x)$ for all $x \in M$ and $g \in G$. Then the pre-variance is well defined in the quotient space by $[x] \mapsto F(x)$.

It is not surprising to use a discrepancy measure which is not a distance, for instance the Kullback-Leibler divergence [KL51] is not symmetric although it is commonly used.

In the proof of inconsistency of Theorem 2.4, we used that the action was isometric in order to simplify the expansion of the variance in Equation (6):

$$\begin{aligned} F(m) &= \mathbb{E}\left(\inf_{g\in G} \|m - g\cdot Y\|^2\right) \\ &= \mathbb{E}\left(\inf_{g\in G} \left[\|m\|^2 - \langle m, g\cdot Y\rangle + \|g\cdot Y\|^2\right]\right), \end{aligned}$$

with $\|g \cdot Y\|^2 = \|Y\|^2$ there was only one term which depends on g: $\langle g \cdot m, Y\rangle$ and the two other terms could be pulled out of the infimum. When the action is no longer isometric we cannot do this trick anymore. To remedy this situation, in this article, we require that the orbit of the template is a bounded set.

In the following, we prove inconsistency even with non isometric action (but only when the noise level is large enough if the template is not a fixed point). The sketches of the different proofs are always the same: finding a point m such that $F(m) < F(t_0)$, in order to do that it suffices to find an upper bound of $F(m)$ and a lower bound of $F(t_0)$ and to compare these two bounds.

3.3 Non Invariant Group Action, with a Subgroup Acting Isometrically

In this subsection G acts on M an Hilbert space. We assume that there exists a subgroup $H \subset G$ such that H acts isometrically on M. As H is included in G, we deduce a useful link between the variance of Y projected in M/H and the pre-variance of Y projected in M/G:

$$F(m) = \mathbb{E}(\inf_{g\in G}\|g\cdot m - Y\|^2) \leq \mathbb{E}(\inf_{h\in H}\|h\cdot m - Y\|^2) = F_H(m).$$

The orbit of a point m under the group action G is $[m] = \{g\cdot m, g \in G\}$, whereas the orbit of the point m under the group action H is $[m]_H = \{h\cdot m, h \in H\}$. Moreover, we call F_H the variance of $[Y]_H$ in the quotient space M/H, and F the variance of $[Y]$ in the quotient space M/G.

3.3.1 Inconsistency when the Template Is a Fixed Point

We begin by assuming that the template t_0 is a fixed point under the action of G:

Proposition 3.3. *Suppose that t_0 is a fixed point under the group action G. Let ϵ be a standardized noise which support is not included in the fixed points under the group action of H, and $Y = \Phi\cdot t_0 + \sigma\epsilon = t_0 + \sigma\epsilon$. Then t_0 is not a minimum of the pre-variance F.*

Proof. We have:

1. Thanks to Corollary 2.5 of Section 2.2 we know that $[t_0]_H = [\mathbb{E}(Y)]_H$ is not the Fréchet mean of $[Y]_H$ the projection of Y into M/H: we can find $m \in M$ such that:

$$F_H(m) < F_H(t_0). \tag{15}$$

   Note that in order to apply Corollary 2.5, we do not need that Φ is included in H, because t_0 is a fixed point.

2. Because we take the infimum over more elements we have:

$$F(m) \leq F_H(m). \tag{16}$$

3. As t_0 is a fixed point under the action of G and under the action of H:

$$F_H(t_0) = F(t_0) = \mathbb{E}(\|t_0 - Y\|^2). \tag{17}$$

With Equations (15)–(17), we conclude that t_0 does not minimize F.

3.3.2 Inconsistency in the General Case for the Template

The following Proposition 3.4 tells us that when σ is large enough then there is an inconsistency.

Proposition 3.4. *We suppose that the template is not a fixed point and that its orbit under the group G is bounded. We consider $A \geq \sup_{g \in G} \frac{\|g \cdot t_0\|}{\|t_0\|}$ and $a \leq \inf_{g \in G} \frac{\|g \cdot t_0\|}{\|t_0\|}$, note that $a \leq 1 \leq A$ and we have:*

$$\forall g \in G \quad a\|t_0\| \leq \|g \cdot t_0\| \leq A\|t_0\|.$$

We note:

$$\theta(t_0) = \frac{1}{\|t_0\|}\mathbb{E}(\sup_{g \in G} \langle g \cdot t_0, \epsilon \rangle) \text{ and } \theta_H = \frac{1}{\|t_0\|}\mathbb{E}\left(\sup_{h \in H} \langle h \cdot t_0, \epsilon \rangle\right).$$

We suppose that $\theta_H > 0$. If σ is bigger than a critical noise level noted σ_c defined as:

$$\sigma_c = \frac{\|t_0\|}{\theta_H}\left[\left(\frac{\theta(t_0)}{\theta_H} + A\right) + \sqrt{\left(\frac{\theta(t_0)}{\theta_H} + A\right)^2 + A^2 - a^2}\right]. \tag{18}$$

Then we have inconsistency.

Note that in Section 2.2 we have proved inconsistency in the isometric case as soon as $\sigma > \frac{2\|t_0\|}{K}$, where $K \geq \theta_H$, then we find in this theorem an analogical sufficient condition on σ where $\left[\left(\frac{\theta(t_0)}{\theta_H} + A\right) + \sqrt{\left(\frac{\theta(t_0)}{\theta_H} + A\right)^2 + A^2 - a^2}\right]$ is a corrective term due to the non invariant action.

We have shown in [DATP17] that if the orbit of the template $[t_0]_H$ is a manifold, then $\theta_H > 0$ as soon as the support of ϵ is not included in $T_{t_0}[t_0]^{\perp}$ (the normal space of the orbit of the template t_0 at the point t_0).

If $[t_0]$ is not a manifold, we have also seen in [DATP17] that $\theta_H > 0$ as soon as t_0 is an accumulation point of $[t_0]_H$ and the support of ϵ contains a ball $B(0,r)$. Hence, $\theta_H > 0$ is a rather generic condition. Condition (18) can be reformulated as follows: as soon as the signal to noise ratio $\frac{\|t_0\|}{\sigma}$ is sufficiently small:

$$\frac{\|t_0\|}{\sigma} < \frac{\theta_H}{\left(\frac{\theta(t_0)}{\theta_H} + A\right) + \sqrt{\left(\frac{\theta(t_0)}{\theta_H} + A\right)^2 + A^2 - a^2}},$$

then there is inconsistency.

We remark the presence of the constants $\theta(t_0)$ and θ_H in Proposition 3.4. This kind of constants were already here in the isometric case under the form $\theta(\frac{t_0}{\|t_0\|}) = \frac{1}{\|t_0\|}\mathbb{E}(\sup_{g\in G}\langle t_0, g\cdot\epsilon\rangle)$, due to the polarization identity (2), we can state that it measures how much the template looks like to the noise after registration, but only in the isometric case. However we can intuit that this constant plays a analogical role in the non isometric case.

Example 3.5. *Let G acting on M, we suppose that G contains $H = O(M)$ the orthogonal group of M. Assume that G can modifying the norm of the template by multiplying its norm by at most 2. Then we can set up $A = 2$ and $a = 0$. By aligning ϵ and $\|t_0\|$ we have $\theta_H = \mathbb{E}(\|\epsilon\|) > 0$, and $\theta(t_0) = A\mathbb{E}(\|\epsilon\|)$ then when the signal to noise ratio $\frac{\|t_0\|}{\sigma}$ is smaller that $\frac{\mathbb{E}(\|\epsilon\|)}{4+\sqrt{20}}$ then there is inconsistency. By Cauchy-Schwarz inequality we have $\mathbb{E}(\|\epsilon\|) \leq \mathbb{E}(\|\epsilon\|^2) = 1$, thus the signal to noise ratio has to be rather small in order to fulfill this condition.*

3.3.3 Proof of Proposition 3.4

We define the following values:

$$\lambda_H = \frac{1}{\|t_0\|^2}\mathbb{E}\left(\sup_{h\in H}\langle h\cdot t_0, Y\rangle\right) \text{ and } \lambda(t_0) = \frac{1}{\|t_0\|^2}\mathbb{E}\left(\sup_{g\in G}\langle g\cdot t_0, Y\rangle\right).$$

Note that λ_H and $\lambda(t_0)$ are registration scores which definitions are the same than the registration score used in the proof of Theorem 2.4 in Section 2 (only the normalization by $\|t_0\|$ is different). The proof of Proposition 3.4 is based on the following Lemma:

Lemma 3.6. *If:*

$$\lambda_H \geq 0, \tag{19}$$

$$a^2 - 2\lambda(t_0) + \lambda_H^2 > 0, \tag{20}$$

then t_0 is not a minimizer of the pre-variance of $[Y]$ in M/G.

How condition (20) can be understood? In order to answer to that question, let us imagine that $G = H$ acts isometrically, then a can be set up to 1, and $\lambda(t_0) = \lambda_H$ the condition (20) becomes $\lambda_H^2 - 2\lambda_H + 1 = (\lambda_H - 1)^2 > 0$ and the conditions of Theorem 4.2 of [DATP17] aimed to

ensure that $\lambda_H > 1$. Now let us return to the non invariant case: if H is strictly included in G such that a is closed enough to 1 and $\lambda(t_0)$ closed enough to λ_H, then on can think that condition (20) still holds. However, the *closed enough* seems hard to be quantified.

Proof of Lemma 3.6. The proof is based on the following points:

1. $F(\lambda_H t_0) \leq F_H(\lambda_H t_0)$,
2. $F_H(\lambda_H t_0) < F(t_0)$.

With items 1 and 2 we get that $F(\lambda_H t_0) < F(t_0)$. Item 1 is just based on the fact that in the map F, we take the infimum on a larger set than on F_H. We now prove item 2, in order to do that we expand the two quantities, firstly:

$$F_H(\lambda_H t_0) = \mathbb{E}\left(\inf_{h\in H} \|h\cdot\lambda_H t_0\|^2 + \|Y\|^2 - 2\langle h\cdot\lambda_H t_0, Y\rangle\right) \tag{21}$$

$$= \lambda_H^2\|t_0\|^2 + \mathbb{E}(\|Y\|^2) - 2\lambda_H\mathbb{E}\left(\sup_{h\in H}\langle h\cdot t_0, Y\rangle\right) \tag{22}$$

$$= \mathbb{E}(\|Y\|^2) - \lambda_H^2\|t_0\|^2,$$

We use the fact that H acts isometrically between Equations (21) and (22) and the fact that $\lambda_H \geq 0$ because $\inf_{a\in A} -\lambda a = -\lambda \sup_{a\in A} a$ is true for any A subset of $\mathbb{R}$ if $\lambda \geq 0$. Secondly:

$$\begin{aligned} F(t_0) &= \mathbb{E}\left(\inf_{g\in G}\|g\cdot t_0\|^2 + \|Y\|^2 - 2\langle g\cdot t_0, Y\rangle\right) \\ &\geq a^2\|t_0\|^2 + \mathbb{E}(\|Y\|^2) - 2\mathbb{E}\left(\sup_{g\in G}\langle g\cdot t_0, Y\rangle\right) \\ &\geq a^2\|t_0\|^2 + \mathbb{E}(\|Y\|^2) - 2\lambda(t_0)\|t_0\|^2 \end{aligned}$$

Then:

$$F(t_0) - F_H(\lambda_H t_0) \geq \|t_0\|^2\left[a^2 - 2\lambda(t_0) + \lambda_H^2\right] > 0,$$

thanks to hypothesis (20).

Proof of Proposition 3.4. In order to prove Proposition 3.4, all we have to do is proving $\lambda_H \geq 0$ and proving that Condition (20) is fulfilled when $\sigma > \sigma_c$. Firstly, thanks to Cauchy-Schwarz inequality, we have:

$$\begin{aligned} \lambda_H &= \frac{1}{\|t_0\|^2}\mathbb{E}\left(\sup_{h\in H}\langle h\cdot t_0, \Phi\cdot t_0 + \sigma\epsilon\rangle\right) \\ &\geq \frac{1}{\|t_0\|^2}\left[-A\|t_0\|^2 + \mathbb{E}(\sup_{h\in H}\langle h\cdot t_0, \sigma\epsilon\rangle)\right] \geq -A + \sigma\frac{\theta_H}{\|t_0\|} \end{aligned}$$

Note that as $\sigma > \sigma_c \geq A\frac{\|t_0\|}{\theta_H}$ we get $\lambda_H \geq 0$, this proves (19). We also have:

$$\begin{aligned} \lambda(t_0) &= \frac{1}{\|t_0\|^2}\mathbb{E}\left(\sup_{g\in G}\langle g\cdot t_0, \Phi\cdot t_0 + \sigma\epsilon\rangle\right) \\ &\leq \frac{1}{\|t_0\|^2}\left[A^2\|t_0\|^2 + \sigma\mathbb{E}\left(\sup_{g\in G}\langle g\cdot t_0, \epsilon\rangle\right)\right] \leq A^2 + \sigma\frac{\theta(t_0)}{\|t_0\|}, \end{aligned}$$

Then we can find a lower bound of $a^2 - 2\lambda(t_0) + \lambda_H^2$:

$$\begin{aligned} a^2 - 2\lambda(t_0) + \lambda_H^2 &\geq a - 2\left(A^2 + \sigma\frac{\theta(t_0)}{\|t_0\|}\right) + \left(\frac{\sigma\theta_H}{\|t_0\|} - A\right)^2 \\ &\geq a^2 - A^2 - 2\frac{\sigma\theta_H}{\|t_0\|}\left(\frac{\theta(t_0)}{\theta_H} + A\right) + \left(\frac{\sigma\theta_H}{\|t_0\|}\right)^2 := P(\sigma) \end{aligned}$$

For $\sigma > \sigma_c$ where σ_c is the biggest solution of the quadratic Equation $P(\sigma) = 0$, we get $a^2 - 2\lambda(t_0) + \lambda_H^2 > 0$ and template estimation is inconsistent thanks to Lemma 3.6. The critical σ_c is exactly the one given by Proposition 3.4.

3.4 Linear Action

The result of the previous part has a drawback, it requires that the group of deformations contains a non trivial subgroup which acts isometrically. We know remove this hypothesis, but we require that the group acts linearly on data.

3.4.1 Inconsistency

In this Subsection we suppose that the group G acts linearly on M. Once again, we can give a criteria on the noise level which leads to inconsistency:

Proposition 3.7. *We suppose that the orbit of the template is bounded with:*

$$\exists a \geq 0, A > 0 \text{ such that } \forall g \in G \quad a\|t_0\| \leq \|g \cdot t_0\| \leq A\|t_0\|.$$

We suppose that $A < \sqrt{2}$. In other words, the deformation of the template can multiply the norm of the template by less than $\sqrt{2}$. We also suppose that:

$$\theta(t_0) = \frac{1}{\|t_0\|}\mathbb{E}\left(\sup_{g\in G}\langle g \cdot t_0, \epsilon\rangle\right) > 0. \tag{23}$$

There is inconsistency as soon as

$$\sigma \geq \sigma_c = \frac{\|t_0\|}{\theta(t_0)}\left[A^2 + \frac{1 + \sqrt{1 - a^2(2 - A^2)}}{2 - A^2}\right].$$

Example 3.8. *For instance if $A \leq 1.2$, then there is inconsistency if $\sigma \geq 7\frac{\|t_0\|}{\theta(t_0)}$.*

Once again we find a condition which is similar to the isometric case, but due to the non invariant action we have here a corrective term which depends on A and a. Note that as G does not act isometrically, results in [DATP17] do not apply in order to fulfill Condition (23). However it is easy to fulfill this Condition thanks to the following Proposition:

Proposition 3.9. *If t_0 is not a fixed point, and if the support of ϵ contains a ball $B(0, \rho)$ for $\rho > 0$ then*

$$\theta(t_0) = \frac{1}{\|t_0\|}\mathbb{E}\left(\sup_{g\in G}\langle g \cdot t_0, \epsilon\rangle\right) > 0.$$

Remark 3.10. *It is possible to remove the condition $A < \sqrt{2}$ in Proposition 3.7. Indeed Let be $h \in G$ such that:*

$$\frac{\sup_{g \in G} \|g \cdot t_0\|}{\|h \cdot t_0\|} < \sqrt{2}.$$

The template t_0 can be replaced by $h \cdot t_0$ since $\Phi t_0 + \sigma\epsilon$ is equal to $\Phi h^{-1} \cdot h t_0$ and applying Proposition 3.7 to the new template $h \cdot t_0$. We get that $h \cdot t_0$ does not minimize the variance F with $A \leq \sqrt{2}$ (because the new template is $h \cdot t_0$). Since $h \cdot t_0$ does not minimize F, the original template t_0 does not minimize the pre-variance F neither, since $F(t_0) = F(h \cdot t_0)$.

This changes the critical σ_c since we apply Proposition 3.7 to $h \cdot t_0$ instead of t_0 itself.

3.4.2 Proofs of Proposition 3.7 and Proposition 3.9

As in Section 3.3 we first prove a Lemma:

Lemma 3.11. *We define:*

$$\lambda(t_0) = \frac{1}{\|t_0\|^2} \mathbb{E}\left(\sup_{g \in G} \langle g \cdot t_0, Y \rangle \right).$$

Suppose that $\lambda(t_0) \geq 0$ and that:

$$a^2 - 2\lambda(t_0) + \lambda(t_0)^2(2 - A^2) > 0. \tag{24}$$

Then t_0 is not a minimum of F.

Proof of Lemma 3.11. Since

$$\forall g \in G \quad a\|t_0\| \leq \|g \cdot t_0\| \leq A\|t_0\|, \tag{25}$$

then by linearity of the action we get:

$$\forall g \in G,\ \mu \in \mathbb{R} \quad a\|\mu t_0\| \leq \|g \cdot \mu t_0\| \leq A\|\mu t_0\|. \tag{26}$$

We remind that:

$$F(m) = \mathbb{E}\left(\inf_{g \in G} \|g \cdot m\|^2 - 2\langle g \cdot m, Y \rangle + \|Y\|^2 \right).$$

By using Equations (25) and (26) we get:

$$F(t_0) \geq a^2\|t_0\|^2 - 2\lambda(t_0)\|t_0\|^2 + \mathbb{E}(\|Y\|^2),$$

We get:

$$\begin{aligned} F(\lambda(t_0)t_0) &\leq \mathbb{E}\left(A^2\|\lambda(t_0)t_0\|^2 + \|Y\|^2 + \inf_{g \in G} - 2\lambda(t_0)\langle g \cdot t_0, Y \rangle \right) \quad (27) \\ &\leq A^2\lambda(t_0)^2\|t_0\|^2 + \mathbb{E}(\|Y\|^2) - 2\lambda(t_0)^2\|t_0\|^2. \end{aligned}$$

Note that we use the fact that the action is linear in Equation (27). We obtain that t_0 is not the minimum of the F:

$$F(t_0) - F(\lambda(t_0)t_0) \geq \|t_0\|^2 \left[a^2 - 2\lambda(t_0) + \lambda(t_0)^2(2 - A^2) \right] > 0.$$

Proof of Proposition 3.7. By solving the following quadratic inequality we remark that:

$$a^2 - 2\lambda(t_0) + (2 - A^2)\lambda(t_0)^2 > 0 \text{ if } \lambda(t_0) > \frac{1 + \sqrt{1 - a^2(2 - A^2)}}{2 - A^2},$$

Besides, as in section 3.3.2 we can take a lower bound of $\lambda(t_0)$ by decomposing $Y = \Phi \cdot t_0 + \sigma\epsilon$ and applying Cauchy-Schwarz inequality $\langle \Phi \cdot t_0, g \cdot t_0 \rangle \geq -A^2 \|t_0\|^2$, we get:

$$\lambda(t_0) \geq -A^2 + \frac{\sigma}{\|t_0\|}\theta(t_0). \tag{28}$$

Thanks to Condition (28) and the fact that $\sigma > \sigma_c$ we get:

$$\lambda(t_0) \geq -A^2 + \frac{\sigma}{\|t_0\|}\theta(t_0) > \frac{1 + \sqrt{1 - a^2(2 - A^2)}}{2 - A^2}$$

Then $\lambda(t_0) \geq 0$ and Condition (24) is fulfilled. Thus, there is inconsistency, according to Lemma 3.11.

Proof of Proposition 3.9. First we notice that:

$$\|t_0\|\theta(t_0) = \mathbb{E}\left(\sup_{g \in G} \langle g \cdot t_0, \epsilon \rangle\right) \geq \mathbb{E}(\langle t_0, \epsilon \rangle) = \langle t_0, \mathbb{E}(\epsilon) \rangle = 0. \tag{29}$$

In order to have $\theta(t_0) > 0$, first we show that it exists $x \in B(0, \rho)$ and $g_0 \in G$ such that

$$\sup_{g \in G} \langle g \cdot t_0, x \rangle \geq \langle g_0 \cdot t_0, x \rangle > \langle t_0, x \rangle .$$

Let $g_0 \in G$ such that $g_0 \cdot t_0 \neq t_0$. There are three cases to be distinguished (see fig. 9):

1. The vectors $g_0 \cdot t_0$ and t_0 are linearly independent. In this case $t_0^\perp \not\subset (g_0 \cdot t_0)^\perp$, then we can find $x \in t_0^\perp$ and $x \notin (g \cdot t_0)^\perp$. Then $\langle t_0, x \rangle = 0$ and $\langle g \cdot t_0, x \rangle \neq 0$, without loss of generality we can assume that $\langle g \cdot t_0, x \rangle > 0$ (replacing x by $-x$ if necessary). We also can assume that $x \in B(0, \rho)$ (replacing x by $\frac{x\rho}{2\|x\|}$ if necessary. Then we have $x \in B(0, \rho)$ and:
$$\langle g_0 \cdot t_0, x \rangle > 0 = \langle t_0, x \rangle .$$
2. If $g_0 \cdot t_0 = wt_0$ with $w > 1$, we take $x = \frac{\rho}{2\|t_0\|} t_0 \in B(0, \rho)$ and we have:
$$\langle g \cdot t_0, x \rangle = w\frac{\rho}{2}\|t_0\| > \frac{\rho}{2}\|t_0\| = \langle t_0, x \rangle .$$
3. If $g_0 \cdot t_0 = wt_0$ with $w < 1$ we take $x = -\frac{\rho}{2\|t_0\|} t_0 \in B(0, \rho)$ and we have:
$$\langle g_0 \cdot t_0, x \rangle = -w\frac{\rho}{2}\|t_0\| > -\frac{\rho}{2}\|t_0\| = \langle t_0, x \rangle .$$

In all these cases we can find x such that $\langle g_0 \cdot t_0, x \rangle > \langle t_0, x \rangle$ By continuity it exists $r > 0$ such that for all y on this ball we have $\langle g \cdot t_0, y \rangle > \langle t_0, y \rangle$. Then the event $\{\sup_{g \in G} \langle g \cdot t_0, \epsilon \rangle > \langle t_0, \epsilon \rangle\}$ has non zero probability, since x is in the support of ϵ we have $\mathbb{P}(\epsilon \in B(x, r)) > 0$. Thus Inequality in (29) will be strict. This proves that $\theta(t_0) > 0$.

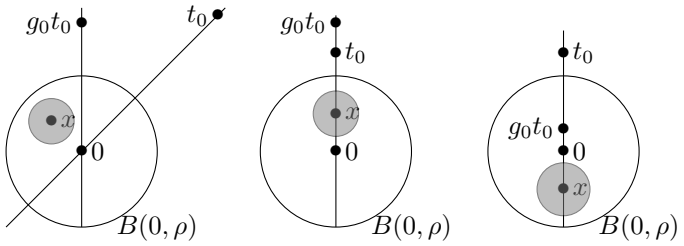


Figure 9: Representation of the three cases, on each we can find an x in the support of the noise such as $\langle x, g_0 \cdot t_0\rangle > \langle x, t_0\rangle$ and by continuity of the dot product $\langle \epsilon, g_0 \cdot t_0\rangle > \langle \epsilon, t_0\rangle$ with is an event with a non zero probability, (for instance the ball in gray). This is enough in order to show that $\theta(t_0) > 0$. (**a**) Case 1: t_0 and $g \cdot t_0$ are linearly independent; (**b**) Case 2: $g \cdot t_0$ is proportional to t_0 with a factor > 1; (**c**) Case 3: $g \cdot t_0$ is proportional to t_0 with a factor < 1.

3.5 Example of a Template Estimation Which is Consistent

In order to underline the importance of the hypotheses, we give an example where the method is consistent:

Example 3.12. *Let M be an Hilbert space and V a closed sub-linear space of M. Then $G = V$ acts on M by (see fig. 10):*

$$(v, m) \in G \times M \mapsto m + v.$$

This action is not isometric, indeed $m \mapsto m + v$ is not linear (except if $v = 0$). However this action is invariant, let us consider $V^\perp$ the orthogonal space of V. The variance in the quotient space is:

$$F(m) = \mathbb{E}\left(\inf_{v\in V} \|m + v - Y\|^2\right) = \mathbb{E}(\|p(m) - p(Y)\|^2) = \mathbb{E}(\|p(m) - p(t_0) + \epsilon\|^2),$$

where $p : M \to V^\perp$ the orthogonal projection on $V^\perp$. Then it is clear that t_0 minimizes F. In fact, $s : [m] \mapsto p(m)$ is just a congruent section of the quotient (see Section 2.1). Here, once again, we see the role played by the the congruent section (when it exists) in order to study the consistency.

Hence, is there a contradiction with Proposition 3.4 or Proposition 3.7 which prove inconsistency as soon as the noise level is large enough? In Proposition 3.4, we require that there is a subgroup acting isometrically, in this example the only element which acts linearly is the identity element $m \mapsto m + 0$, then $H = \{0\}$ is the only possibility, however the support of the noise should not be included in the set of fixed point under the group action of H. Here, all points are fixed under H, hence it is not possible to fulfill this condition. Example 3.12 is not a contradiction with Proposition 3.4, it is also not a contradiction with Proposition 3.7 since it does not act linearly on data.

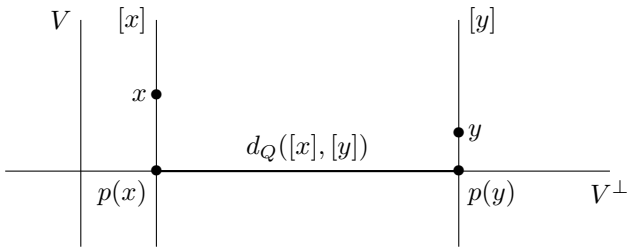


Figure 10: In the case of affine translation by vectors of V, the orbits are affine subspace parallel to V. The distance between two orbits $[x]$ and $[y]$ is given by the distance between the orthogonal projection of x and y in $V^\perp$. This is an example where template estimation is consistent.

3.6 Inconsistency with Non Invariant Action and Regularization

In practice people add a regularization term in the function they minimize in LDDMM [BMTY05, DPC$^+$14], or in Demons [LGP$^+$13] etc. Because, if one considers two points, one does not want necessarily to fit one with the other. Indeed, even if one deformation matches exactly these two points, it may be an unrealistic deformation. So far, we did not study the use of such a term in the inconsistency.

3.6.1 Case of Deformations Closed to the Identity Element of G

If we suppose that the deformations Φ of the template is closed to identity, it is useless to take the infimum over G because G contains big deformations. Perhaps one of these big deformations can reaches the infimum in F, but this element is not the one which deformed the template in the generative model. Then such big deformations should not be taken into account. That is why, if we suppose that G can be equipped with a distance d_G, then we can assume that there exists $r > 0$ such that the deformation Φ belongs almost surely to

$$\mathcal{B} = B(e, r) = \{g \in G, \quad d_G(e, g) < r\}.$$

Instead of defining $F(m)$ as $\mathbb{E}(\inf_{g\in G} \|g \cdot m - Y\|^2)$, one can define $F(m) = \mathbb{E}(\inf_{g\in \mathcal{B}} \|g \cdot m - Y\|^2)$, and the previous proofs will still be true, when replacing for instance $\lambda(t_0)$ by $\lambda(t_0) = \frac{1}{\|t_0\|^2}\mathbb{E}(\sup_{g\in\mathcal{B}} \langle g \cdot t_0, Y\rangle)$ etc. Likewise we need to replace the hypothesis "the support of ϵ is not included in the set of fixed points " by "the support of ϵ in not included is the set of fixed points under the action restricted to $\mathcal{B}$".

Note that restraining ourselves to $\mathcal{B}$ is equivalent to add a following regularization on the function F:

$$F(m) = \mathbb{E}\left(\inf_{g\in G} \|g \cdot m - Y\|^2 + Reg(g)\right) \text{ with } Reg(g) = \begin{cases} 0 & \text{if} \quad g \in \mathcal{B} \\ +\infty & \text{if} \quad g \notin \mathcal{B} \end{cases}.$$

Moreover considering only the elements in $\mathcal{B}$ will automatically satisfy the condition $A < \sqrt{2}$ in Proposition 3.7 as long as the group G acts continuously on the template, if r is small enough.

3.6.2 Inconsistency in the Case of a Group Acting Linearly with a Bounded Regularization

In this Section we suppose that the group G acts linearly. We also suppose that $A < \sqrt{2}$. The regularization term is a bounded map $Reg : G \to [0, \Omega]$. With this framework, we still able to prove that there is inconsistency as soon as the noise level is large enough:

Proposition 3.13. *Let G be a group acting linearly on M. We suppose that the orbit of the template t_0 is bounded with $A = \sup_{g \in G} \frac{\|g \cdot t_0\|}{\|t_0\|} < \sqrt{2}$, the generative model is still $Y = \Phi \cdot t_0 + \sigma\epsilon$. We define the pre-variance as:*

$$F(m) = \mathbb{E}\left(\inf_{g \in G} \left(\|Y - g \cdot m\|^2 + Reg(g) \right) \right).$$

Then as soon as the noise level is large enough, i.e.,:

$$\sigma > \sigma_c = \frac{\|t_0\|}{\theta(t_0)} \left[A^2 + \frac{1 + \sqrt{1 - (a^2 + \frac{\Omega}{\|t_0\|^2})(2 - A^2)}}{2 - A^2} \right].$$

Then t_0 is not a minimizser of F.

The proof is exactly the same as the Proof of Proposition 3.7, we take 0 as a lower bound of the the regularization term in the lower bound of $F(t_0)$, and we take Ω as a upper bound of the regularization term in the upper bound of $F(\lambda(t_0)t_0)$. We solve a similar quadratic equation in order to find the critical σ.

4 Conclusions and Discussion

We provided an asymptotic behavior of the consistency bias when the noise level σ tends to infinity in the case of isometric action. As a consequence, the inconsistency can not be neglected when σ is large. When the action is no longer isometric, inconsistency has been also shown when the noise level is large.

However, we have not answered this question: can the inconsistency be neglected? When the noise level is small enough, then the consistency bias is small [MHP16, DATP17], hence it can be neglected. Note that the quotient space is not a manifold, this prevents us to use a priori the Central Limit theorem for manifold proved in [BB08]. However, if the Central Limit theorem could be applied to quotient space, the fluctuations induces an error which would be approximately equal to $\frac{\sigma}{\sqrt{I}}$ and if $K \ll \frac{1}{\sqrt{I}}$, then the inconsistency could be neglected because it is small compared to fluctuation. One way to avoid the inconsistency is to use another framework, for a instance a Bayesian paradigm [CDH16].

In the numerical experiments we presented, we have seen that the estimated template is more crispy that the true template. The intuition is that the estimated template in computational anatomy with a group of diffeomorphisms is also more detailed. However, the true template is almost always unknown. It is then possible that one think that the computation of the template succeeded to capture small details of the template while it is just an artifact due to the inconsistency. Moreover in order to tackle this question, one needs to have a good modeliation of the noise, for instance in [KSW11], the observations are curves, what is a relevant noise in the space of curves?

In this article, we have considered actions which do not let the distance invariant. Although we have only shown the inconsistency as soon as the noise level is large enough, the inequality used where not optimal at all, surely future works could improve this work and prove that inconsistency appears for small noise level. Moreover a quantification of the inconsistency should be established.

References

[AAT07] Stéphanie Allassonnière, Yali Amit, and Alain Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

[BB08] Abhishek Bhattacharya and Rabi Bhattacharya. Statistics on riemannian manifolds: asymptotic distribution and curvature. *Proceedings of the American Mathematical Society*, 136(8):2959–2967, 2008.

[BC11] Jérémie Bigot and Benjamin Charlier. On the consistency of fréchet means in deformable models for curve and image analysis. *Electronic Journal of Statistics*, 5:1054–1089, 2011.

[BMTY05] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

[CDH16] Wen Cheng, Ian L Dryden, and Xianzheng Huang. Bayesian registration of functions and curves. *Bayesian Analysis*, 11(2):447–475, 2016.

[Cha13] Benjamin Charlier. Necessary and sufficient condition for the existence of a fréchet mean on the circle. *ESAIM: Probability and Statistics*, 17:635–649, 2013.

[CMT+04] Timothy F Cootes, Stephen Marsland, Carole J Twining, Kate Smith, and Christopher J Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *European conference on computer vision*, pages 316–327. Springer, 2004.

[CT65] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[DATP17] Loïc Devilliers, Stéphanie Allassonnière, Alain Trouvé, and Xavier Pennec. Template estimation in computational anatomy: Fréchet means in top and quotient spaces are not consistent. *SIAM Journal on Imaging Sciences*, 2017.

[DPA17] Loïc Devilliers, Xavier Pennec, and Stéphanie Allassonnière. Inconsistency of Template Estimation with the Fréchet mean in Quotient Space. In *Information Processing in Medical Imaging 2017*, Boone, United States, June 2017. Martin Styner and Marc Niethammer and Dinggang Shen and Stephen Aylward and Ipek Oguz and Hongtu Zhu.

[DPC+14] Stanley Durrleman, Marcel Prastawa, Nicolas Charon, Julie R Korenberg, Sarang Joshi, Guido Gerig, and Alain Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*, 101:35–49, 2014.

[GM01] CA Glasbey and KV Mardia. A penalized likelihood approach to image warping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):465–492, 2001.

[GMT00] A. Guimond, J. Meunier, and J.-P. Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192–210, 2000.

[Goo91] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 285–339, 1991.

[HCG+13] Sebastian Hitziger, Maureen Clerc, Alexandre Gramfort, Sandrine Saillet, Christian Bénar, and Théodore Papadopoulo. Jitter-adaptive dictionary learning-application to multi-trial neuroelectric signals. *arXiv preprint arXiv:1301.3611*, 2013.

[Huc11] Stephan Huckemann. Inference on 3d procrustes means: Tree bole growth, rank deficient diffusion tensors and perturbation models. *Scandinavian Journal of Statistics*, 38(3):424–446, 2011.

[JDJG04] Sarang Joshi, Brad Davis, Mathieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23:S151–S160, 2004.

[KKD+11] Sebastian Kurtek, Eric Klassen, Zhaohua Ding, Malcolm J Avison, and Anuj Srivastava. Parameterization-invariant shape statistics and probabilistic classification of anatomical surfaces. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 147–158. Springer, 2011.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[KSW11] Sebastian A. Kurtek, Anuj Srivastava, and Wei Wu. Signal estimation under random time-warpings and nonlinear signal alignment. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett,

F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 675–683. Curran Associates, Inc., 2011.

[LGP+13] Herve Lombaert, Leo Grady, Xavier Pennec, Nicholas Ayache, and Farida Cheriet. Spectral log-demons: Diffeomorphic image registration with very large deformations. *International Journal of Computer Vision (IJCV)*, pages 1–18, 2013.

[MHP16] Nina Miolane, Susan Holmes, and Xavier Pennec. Template shape estimation: correcting an asymptotic bias. *arXiv preprint arXiv:1610.01502*, 2016.

[PZ16] Victor M Panaretos and Yoav Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016.

[Roh03] F James Rohlf. Bias and error in estimates of mean shape in geometric morphometrics. *Journal of Human Evolution*, 44(6):665–683, 2003.

[Tho42] Darcy Wentworth Thompson. On growth and form. *On growth and form.*, 1942.

[Zie77] Herbert Ziezold. On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, pages 591–602. Springer, 1977.