Ph.D. Project
Geometric inference for Data analysis: a Geometric Measure Theory perspective

Charly Boricaud

Ph.D. Advisors: Blanche Buet (Univ. Paris-Saclay), Gian Paolo Leonardi (Univ. Trento) and Simon Masnou (Univ. Lyon 1)

Continuous definitions (such as those of surface, regularity, dimension, curvatures ...) generally cannot be readily given a discrete counterpart. Moreover, this discrete counterpart is generally not unique and highly scale-dependent. There are multiple ways of developing a theory for discrete surfaces and the choice of an appropriate framework is directly related to the kind of discrete data we aim to process, and for which purpose i.e. the type of surfaces we try to model. Regarding the kind of discrete data, two different situations occur: either they have been collected in an external context and come in some given form one has to deal with, or one has the freedom to decide which discretization is best suited for this issue. Let us mention some examples of discrete representations: triangulated surfaces, digital shapes, graph representations, level sets and diffuse interfaces etc. In this project, we propose to focus on **point cloud** data (on the discrete side). On the continuous side, the denomination "surface" encompasses a wide variety of objects ranging from usual 2–dimensional surfaces embedded in $\mathbb{R}^3$ to any dimension and co-dimension submanifold, abstract Riemannian and sub-Riemannian manifolds, rectifiable sets, tree-like and graphs structures, stratified spaces etc.

In this project, we propose to focus on **unstructured data** in the sense that we do not have any underlying parametrization or topological information associated with our data, a typical example being **point cloud** data (e.g. obtained from scan acquisition) or different kinds of diffuse approximations (as in MRI for instance). Our motivation is twofold: first, a large range of data-types initially comes without any parametrization information. Moreover, let us point out that the construction of such a parametrization (for instance the definition of a triangulation starting from a point cloud) is a challenging active topic in itself, and a better understanding of unstructured data is an essential pre-processing step. Of course, working with parametrized objects has a lot of practical advantages making them quite popular. Nonetheless, it has also a well-known drawback as it does not allow to handle changes of topology and self-crossings. Furthermore, most of the collected data are far from lying on a nice manifold. In such a case, discrete parametrized surfaces cannot represent them properly and then the problem stands on choosing a manifold model.

Geometric measure theory actually shows a major advantage: both discrete and continuous surfaces can be associated with a natural measure and thus naturally lie in the same space. It is in particular possible to say that a point cloud is close to a surface in the sense that it is close to a measure supported by the surface, with different possible choices of distances between measures to quantifiy the closeness. In geometric inference, such a closeness in terms of measures is a classical asumption in order to establish convergence of geometric estimators (tangent, curvature, second fundamental form, Laplace-Beltrami operator, see for instance [CSM06, BLM17, TGHS18, BLM19, BR20]). However, such an asumption is not meaningless and essentially implies that the set of point is uniformly distributed along the underlying continuous surface (which is rarely true) or that it is possible to weight points to rectifiy the sampling.

In most cases, we are only provided with sets of points in $\mathbb{R}^n$ (for some n) and we need to infer weights and approximate tangents (not to mention dimension) in order to infer the varifold structure from the data. We intend to explore the statistical aspects of deterministic curvature estimators developed in [BLM17] in the varifold framework, in the spirit of [AL19]. As those curvature estimators rely on varifold structure and convergence has been established with respect to weak star topology of varifolds, it is necessary to investigate varifold reconstruction from the data in this statistical framework. It is a long-time standing question that has been given multiple relevant answers. Yet, to the best of our knowledge, there is no mass estimator proven to be convergent, considering integral varifolds for instance.

Before going into more details about the objectives of this project, we would like to underline that we intend to **implement any estimator** we would investigate, using **Python programming language**. Our motivation is twofold: first, it is crucial to observe how a method perform from a practical standpoint. It may help understanding theoretical features of a given estimator such as resilience to noise, potentially depnding on the nature of the noise, robustness to outliers, irregular sampling, it may also indicate for which data it is better suited (depending on smoothness, high/low dimension and codimension, topology, boundaries, self-intersections, branching/stratified structures...). We point out that almost all generated or collected data are given in the form of point clouds, from scan acquisition of 3D objects to sets of images (e.g. medical radiographs of the same body part for a sample of patients) or meteorological data. Furthermore, when some theoretical background has been set up, we consider that it is essential to compare convergence guarantees with numerical results, both on academic data and real-world data.

Statistical framework.

In this framework, we assume that the continuous object S is given through a probability measure μ supported in S and our data are obtained by sampling μ with N points: $(X_1, \dots, X_N) \sim \mu$ is an i.i.d. sample and our data is an instance of the empirical measure

$$\mu_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i} \, .$$

The specific and important case where S is d–dimensional submanifold of $\mathbb{R}^n$ and μ is the volume form in S (possibly weighted by some density) has been investigated with varying degrees of formality and different assumptions on the regularity of the manifold. In such a context, one can devise and analyse estimators of geometric quantities such as tangent space and second fundamental form, as well as estimators of S itself. A very complete analysis is carried out in [AL19]: the authors establish general minimax bounds for the aforementioned estimators, with rates of convergence involving the size of the sample N, the dimension of the manifold d and its order of regularity. They also evidenced the need of a global assumption: in addition to C^k–regularity, one has to work with submanifolds S sharing a uniform lower bound on their reach to obtain convergence of the geometric estimators in strong topology ($\mathrm{L}^\infty(S)$ for tangent spaces or second fundamental form and Hausdorff distance for S itself).

We intend to consider the problem at hand from the measure/varifold perspective. Let us describe the concrete differences that come along with such a change of standpoint. First, while the smooth manifold setting naturally allows to consider the tangent spaces and the second fundamental form as smooth functions in S, we only have weaker counterparts (measure and distributions) in the varifold setting. Consequently, we do not expect that the associated estimators converge in $\mathrm{L}^\infty(S)$ but rather as measures or almost everywhere in S. Similarly, we do not wish to reconstruct S for Hausdorff distance but rather $\mathcal{H}^d_{|S}$ for distances related to weak convergence of measures (such as flat and Prokhorov distances). Weakening from L^∞–type to L^1–type (even measure–type) the distances we use to quantify the accuracy of estimators, we should escape the stronger necessary conditions established in [AL19].

Pointwise density estimation.

Depending on the nature of the collected data, the measure μ is not necessarily uniformly distributed in S which consequently is also reflected by μ_N. In such a case, it is important to decouple the geometric information contained in S from the whole information encoded by μ. More explicitly, assume that $\mu = p\mathcal{H}^d_{|S}$ for some positive density function p that we want to recover.
In practice, a common approach consists in computing local averages of the following form

$$p_i = \frac{1}{N} \sum_{j=1}^{N} \eta \left(\frac{|x_i - x_j|}{\epsilon} \right) .$$

Reading p_i as the integral of $\eta \left(\frac{|x_i - \cdot|}{\epsilon} \right)$ with respect to μ_N, we can relate p_i to the analogous quantity obtained by switching the integration to μ.
On one hand, the weak convergence of μ_N towards μ holds with probability 1 for very general S and μ (far beyond our euclidean scope), it is known as Glivenko-Cantelli theorem. Moreover, in the fundamental work [D69], such a weak convergence is quantified in terms of bounded Lipschitz distance $\Delta(\mu,\ \mu_N)$ and Prokhorov distance. In particular, the author gives asymptotic rate of convergence of the form

$$\mathbb{E}\left[\Delta(\mu,\ \mu_N)\right] \propto N^{-\frac{1}{k}}$$

provided that the minimal number of balls of radius δ required to cover S is of order δ^{-k}. We expect $k = d$ for a d–Ahlfors measure μ.
On the other hand, the rate of convergence of local averages of p around x (towards $p(x)$) depends on the (local) regularity of S, though the convergence itself holds for a.e. $x \in S$ for a very general set S. Eventually, we should obtain a rate of convergence resulting from the competition between the averaging parameter ϵ and the number of points in the sample N, depending on d and not on the ambient dimension n.

Tangent space and varifold estimation.

The very practical problem of estimating tangent spaces is a long-standing question that has already been substantially addressed. However, there are few theoretical guarantees when the set S is only assumed to be rectifiable. Let us propose at least two possible approaches for such weakly regular sets.

One path adapts the least square strategy: among all $P \in \mathrm{G}_{d,n}$, take the minimizer of some well-suited energy involving the so-called *heightexcess* at x,

$$\mathrm{heightex}_\mu(x, \epsilon, P) = \frac{1}{\epsilon^d} \int_{B_\epsilon(x)} \left(\frac{\mathrm{dist}(y - x, P)}{\epsilon} \right)^2 d\mu(y) .$$

It is shown in [Bue15] that an averaged version $E_{\alpha_i}(P) = \int_{\alpha_N}^1 \mathrm{heightex}_{\mu_N}(x, r, P) \frac{dr}{r}$ Γ–converges to $E_0(P) = \int_0^1 \mathrm{heightex}_\mu(x, r, P) \frac{dr}{r}$ whose unique minimizer is the approximate tangent space $T_x S$, for a.e. $x \in S$, under some technical assumptions. We can infer that minimizers of E_{α_i} tend to $T_x S$, but no speed of convergence follows from the proof of Γ–convergence. Taking into account that the tangent space estimation is not the central question in [Bue15] we believe that averaging on the radius is not needed and we can directly work on the Γ–convergence of $\mathrm{heightex}_{\mu_N}(x, \epsilon_N, \cdot)$. The relation between ϵ_N and N should involve the bounded Lipschitz distance $\Delta(\mu, \mu_N)$. We are then left with the speed of convergence to estimate.

Another path follows the reconstruction work carried out in [Tin19] for an immersed manifold S directly considering the rescaled covariance matrix to approximate $T_x S$:

$$\frac{1}{r^d}\int_{B_\epsilon(x)} \frac{y-x}{r} \otimes \frac{y-x}{r}\, d\mu(y)\ .$$

The authors establish pointwise convergence, with speed depending on ϵ and the bounded Lipschitz distance between μ and μ_N both restricted to $B_\epsilon(x)$ and renormalized to mass 1. A very interesting point of this second approach is that the author of [Tin19] pushed the analysis further to obtain the reconstruction of the whole varifold associated with S, with very explicit convergence rate for the p–Wasserstein distance between the exact and the reconstructed varifolds. Loosening the assumption from immersed manifold to rectifiable, we should be able to obtain a.e. convergence of the tangent space estimator via rescaled covariance. Note that for a rectifiable set, $T_x S$ is only defined a.e. in S so that one cannot hope for a better pointwise convergence.

The reconstruction of the whole varifold then requires to put density and tangent estimation together. It is a very important stage to reach since higher order estimators will rely on the quality of the approximation of the varifold structure. We aim to prove weak convergence of the reconstructed varifold towards the rectifiable varifold and then to quantify such a convergence in terms of bounded Lipschitz and Prokhorov distances (with possible variations around such distances). Considering p–Wasserstein distances would be interesting in this perspective.

Curvatures estimation.

Once the varifold estimation has been investigated, it will be possible to tackle the curvatures estimators established in [BLM17, BLM19] (see also [BR20] for a more explicit rate convergence in terms of Bounded Lipschitz distance Δ and a Prokhorov type distance η_d). The deterministic convergence rate depends on Δ and η_d though not in a straightforward way and the statistical asymptotic rate requires additional analysis. Then, building on the fact that curvatures encode some topological information on the shape through the Gauss-Bonnet theorem, we propose to associate a scale dependent family of genus with a given point cloud (by computing the weighted sum of the approximate Gauss curvature at scale ϵ) and investigate its theoretical and numerical properties. We also consider the issue of devising approximate Laplace-Beltrami operators on point cloud varifolds and extend convergence analysis to the non smooth setting, see for instance [TGHS18].

Noise model.

The issue of handling noise actually intersects each previous task since the sensitivity to noise can vary from one estimator to another. Let us identify two different contributions to be considered when adding a noise term Z_i to each point X_i (for $i = 1 \dots N$). The natural case that is often considered in literature is the case where Z_i is normal to the surface at X_i. Such a noise may directly impact the geometry of the shape and when working with estimations in strong topology such as L^∞ norm for tangent or curvatures and Hausdorff distance for reconstruction of S, one usually has to restrict to uniformly bounded noise, as in [AL19] where the model can handle very general noise up to those 2 assumptions (normal and bounded noise). However, when working in the measure theoretic framework, which is the point of our project, it is reasonable to consider noise that is not uniformly bounded but whose standard deviation is bounded. Of course, relaxing such a constraint is not straightforward, and yet, it would allow to include Gaussian noise for instance. Reasoning in terms of the continuous object, it is natural to discard tangential contribution since tangential motion of S leaves S invariant, when a parametrization is given, tangential motion results in a reparametrization of S. Unfortunately, let us raise two objections. First, at the discrete level, a tangential motion can result in a significant motion outside S, especially in highly curved parts or singular parts of S. Moreover, even at a continuous level, one has to assume some coherence (spatial regularity) of the tangential motion to match it with a reparametrization, otherwise tangential motion can modify the distribution μ while preserving its support S. At the discrete level, points may concentrate in some parts and create small holes in other parts. This issue is closely related to the density estimation (first part of the project) in which we assume that the initial distribution $\nu = \mathrm{cte}\mathcal{H}^d_{|S}$ is uniform in S while we observe the corrupted distribution $\mu = p(x)\nu$: estimating p is a first step so as to understand the loss of uniformity coming with p and try to rectify it.

Let us eventually observe that in the noise free context, there are two important parameters, the number of points N in the sample and the scale $\epsilon > 0$ at which we compare μ and μ_N. On one hand, just following some naive intuition we should take $\epsilon > 0$ large enough with respect to some power of $1/N$ (the number of points in a ball of radius ϵ centred at S is roughly $\epsilon^d \times N^{-1}$) to ensure that our estimator captures geometric information on μ and not artifacts due to our specific μ_N. Loosely speaking, we average μ and μ_N at some scale $\epsilon > 0$ that allows to compare them. On the other hand, ϵ must be chosen small enough to avoid over-smoothing effect, this requirement only involves μ. The result of both is a competition between ϵ and N that can be optimized once the mean speed of convergence of the considered estimator is explicit enough.

References

[AL19] Aamari, E. and Levrard, C. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47:177–204, 2019.

[Bue15] B. Buet. Quantitative conditions of rectifiability for varifolds. *Ann. Institut Fourier*, 2015, `https://arxiv.org/abs/1409.4749v1`.

[BLM17] B. Buet, G. P. Leonardi, and S. Masnou. A varifold approach to surface approximation. *ARMA*, 2017, `https://arxiv.org/abs/1609.03625v1`.

[BLM19] B. Buet, G. P. Leonardi, and S. Masnou. Weak and approximate curvatures of a measure: a varifold perspective. *Arxiv*, 2019, `https://arxiv.org/abs/1904.05930v2`.

[BR20] B. Buet, and M. Rumpf. Mean curvature motion of point cloud varifolds. *Arxiv*, 2020, `https://arxiv.org/pdf/2010.09419.pdf`.

[CSM06] D. Cohen-Steiner and J.-M. Morvan. Stability of Curvature Measures. *J. Differential Geom.*, 74(3):363–394, 2006.

[D69] R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[S83] L. Simon. Lectures on Geometric Measure Theory, 1983.

[Tin19] R. Tinarrage. Recovering the homology of immersed manifolds. *Arxiv*, 2019, `http://arxiv.org/abs/1912.03033`.

[TGHS18] N. Trillos, M. Gerlach, M. Hein, D. Slepcev. Error Estimates for Spectral Convergence of the Graph Laplacian on Random Geometric Graphs Toward the Laplace–Beltrami Operator. *Foundations of Computational Mathematics*, 2018.

[TS15] N. Trillos, D. Slepcev. On the Rate of Convergence of Empirical Measures in ∞–transportation Distance *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.