
Project number IST-25582

**CGL**

Computational Geometric Learning

# D1.2: Work Package 1 [Period 2] Report

**STREP**

**Information Society Technologies**

| | |
|---|---|
| Period covered: | November 1, 2010–October 31, 2011 |
| Date of preparation: | October 30, 2012 |
| Date of revision: | |
| Start date of project: | November 1, 2010 |
| Duration: | 3 years |
| Project coordinator name: | Joachim Giesen (FSU) |
| Project coordinator organisation: | Friedrich-Schiller-Universität Jena<br>Jena, Germany |

In the following we describe the work done within Work Package 1 within the second period. In the description we follow the structure imposed by the tasks for this work package and period. Wherever there is no deviation and all goals have been met, it is *not* mentioned specifically. We start by restating the objectives for Work Package 1 and conclude with a discussion of the milestones for Period 1.

## Objectives

Geometric inference aims at inferring topological or geometric properties of a presumably unknown shape from a set of data points sampling it. A stronger goal is to compute an approximation from the sample points that shares many of the geometric and topological properties with the unknown shape. The latter problem, known as *curve/surface reconstruction* in low dimensions or as *manifold learning* in high dimensions, has attracted strong interest in many scientific fields where the data may often be represented by point sets in high dimensional spaces. Although point sets may live in high dimensional spaces, one often expects them to be located around unknown, possibly non-linear, low dimensional shapes. An important class of shapes is formed by smooth submanifolds of Euclidean spaces, but also more general compact subspaces need to be considered. It is then desirable to infer topological features (dimension, Betti numbers) and geometric characteristics (singularities, volume, curvature) of these shapes from the data. In some cases the shapes of interest may be known through a set of conditions - e. g., solution sets of geometric constraint problems, like boundaries of configuration spaces in robotics or iso-energy hypersurfaces in six-dimensional phase space in physical applications - that allow to sample them. It then desirable to study and design efficient sampling strategies to reliably approximate or reconstruct them. Building on recent results in computational geometry and geometric approximation combined with statistical approaches the main goal of this work package is to provide new algorithmic tools based upon strong theoretical models to infer and compute relevant topological and geometrical properties of the shapes from which the data are drawn.

## Tasks

### Task 1.a: Inference from noisy data

The goal of geometric inference is to extract topological and geometric structure from data to get a high-level informative understanding of these data and of the spaces they originate from. During the first period we have developped an approach based on distance-like functions that provides a framework allowing to deal with data sets embedded in Euclidean spaces of any dimension and possibly corrupted by noise. However data are not always embedded in Euclidean spaces and often come as set of points with pairwise distances information, i.e. metric spaces. During the second period we focused on the problem of inference for general metric spaces. In this case explicit reconstruction of geometric objects is sometime out of reach and even non relevant. The main challenge is then to be able to infer topological and geometric information without the need for a costly explicit reconstruction. For that purpose we have concentrated our efforts on approximation of data by simplicial complexes from which topological information can be robustly infered through the use of persistence homology. More precisely we have initiated and obtained results in the following directions:
- *General theory of persistence for topological inference:* the theory of homology persistence, intro-

duced and formalized by Edelsbrunner et al. a decade ago, provides very powerful tools to infer robust topological information from data at different scales. However the classical theory restricting to tame functions and filtrations of finite simplicial complexes does not allow to directly address important questions encountered in topological data analysis (e.g. multiscale homology inference for metric spaces or scalar fields analysis on discrete data). To overcome this issue we have extended generalized persistence homology theory and its stability results that are of fundamental importance for geometric inference [5]. As an application we have proven the robustness of the persistent homology of various families of geometric filtered complexes built on top of compact metric spaces or spaces endowed with a similarity measure [6]. We are currently working on the extension of these results to families of complexes inspired from the distance-to-measure framework [4].
*Simplicial complexes for inference:* The so-called Čech and Vietoris-Rips filtrations built on top of data sets are useful tools to infer topological information from data. However they are often too large to be constructed in full. To overcome this issue we have shown how to construct an $O(n)$-size filtered simplicial complex on an $n$-point metric space such that the persistence diagram is a good approximation to that of the Vietoris-Rips filtration [14].The filtration can be constructed in $O(n \log n)$ time. The constants depend only on the doubling dimension of the metric space and the desired tightness of the approximation. For the first time, this makes it computationally tractable to approximate the persistence diagram of the Vietoris-Rips filtration across all scales for large data sets. Regarding the Čech complex we have shown that filtering its barycentric decomposition by the cardinality of the vertices captures precisely the topology of $k$-covered regions among a collection of balls for all values of $k$ [15].
- *Metric graph reconstruction:* data sometimes happen to be sampled around some non-manifold 1-dimensional structures that can be represented as metric graphs. This is, for example, the case when one considers GPS traces recorded along a road network, or distribution of earthquake epicenters that are usually concentrated along faults or in astrophysics where the distribution of galaxies in the universe is known to exhibit some filamentary structures. Even when sampled in an Euclidean space, these data usually come with an intrinsic metric that is not the restriction of the Euclidean one. To infer the topology of such structures we have designed a metric graph reconstruction algorithm that take as input a discrete metric space which is Hausdorff close to a metric graph and output a topologically correct and quasi-isometric reconstruction of this graph [1]. Our algorithm being rather sensitive to the quality of the sampling we are currently working on some generalization of our method that will lead to much more robust and efficient algorithms.

### Task 1.b: Clustering

A coreset can be viewed as a small weighted point set that approximates a big input point set with respect to a clustering problem. For example, if we consider $k$-means clustering, i.e. the problem to find $k$ centers such that the sum of squared distances of the input points to their nearest center is minimized, then a (strong) coreset guarantees that for every set of $k$ centers the cost of the coreset deviates by at most a $(1+\epsilon)$-factor from the original point set $P$. Coresets are typically be used to obtain streaming and distributed algorithms.

In [9] we prove the surprising result that strong coresets exist whose cardinality is independent of the dimension of the input space as well as the number of input points. We extend out result to the problem of finding $k$ affine subspaces of dimension $j$. We slightly change the definition of a coreset by allowing that the cost of the coreset plus some fixed additive constant approximates the cost of the original point set (note that this still is a multiplicative approximation). We then prove that by writing the point set as a matrix $A$ (whose rows are the points) and replacing $A$ by

its best rank $k$ approximation $A'$ (which can be efficiently computed using, for example, singular value decomposition), the rows of $A'$ form a point set in a low dimensional space that approximates the original point set with respect to subspace clustering (again, we have to add a constant to approximate the cost). As a special case of this we obtain, that computing $A'$ and solving $k$-means upto a factor of $(1+\epsilon)$ for $A'$, then the result will be a $(1+O(\epsilon)$-approximation for the $k$-means problem with input $A$.

Furthermore, we started to study data structures for clustering range query, i.e. we would like to have a data structure that efficiently returns a $k$-clustering (say, $k$-means clustering) of all points in a given query range. In order to support efficient clustering queries, we observe that in many data structures that support range queries, the input point set is subdivided into certain subsets and the range query is answered by combining these subsets. Now, one can replace each of these subsets by a coreset and instead of the union of the subsets return the union of the corresponding coresets. Then one can solve the clustering problem on the coreset. Since the coresets are typically very small, one can efficiently answer range clustering queries in this way.

Deviations: TUDO has spend 10.71 funded PM (instead of 10PM) on clustering (Task 1.b). It is planned to compensate for this by underspending in the third project period.

### Task 2.a: Scale selection

It appears that data often carries different topological and geometric structures at different scales, making the automatic scale selection problem from a statistical point of view particularly difficult in its whole generality. After the results on scale selection for simplicial complexes approximation obtained during Period 1, we have focused our efforts to statistical approaches for homology and persistent homology that allow multiscale inference (persistent diagrams can be seen as summaries of the topological/homological structure of the data at all scales).

As mentioned in Task 1.a, thanks to our results on general persistence, we have proven stability results for the persistence diagrams of various filtrations built on top of metric spaces and spaces endowed with a similarity measure [6]. It turns out that these results provide powerful tools to study the statistical behavior of the persistence diagrams obtained from data randomly sampled on these spaces. We are currently working on this approach and results are expected during Period 3.

We have also considered the statistical problem of estimating the homology groups of a manifold from noisy samples under several different noise models [3]. We have derived upper and lower bounds on the minimax risk for this problem. Our upper bounds are based on estimators which are constructed from a union of balls of appropriate radius around carefully selected points. In each case we have established complementary lower bounds using Le Cam's lemma. We expect these results will extend to the persistence diagrams of distance functions to manifolds or more generally to the distance to measures framework introduced during Period 1 [4].

### Task 2.b: Morse-Smale and flow complexes

The rationale here was *to investigate how to build Morse-Smale complexes for scalar functions known only through a finite set of sample points*, an endeavor raising two types of difficulties. The first one is theoretical, since the correctness (geometric and topological) of the structure computed should ideally be assessed. That is, following the type of analysis developed for the surface reconstruction problem, one would like to derive sufficient conditions on the sampling so that correctness holds. The second one is practical, since for selected examples where the ground truth i.e. the correct M-S complex is known, a comparison of the diagram computed against this ground truth is in order.

From the theoretical standpoint, while this report is being written, we are finalizing a paper entitled *A Heuristic Construction of Discrete Morse-Smale Diagrams based on Samplings and Bifurcations Diagrams*. As the name suggest, nearest neighbor graphs (NNG) are used to defined (pseudo-)gradient vector fields, from which critical points and (uns-)stable manifolds are computed. We emphasize that this construction can be carried out in any dimension since only NNG are used. We plan to submit this paper to ACM SoCG 2013.

From the practical standpoint, in collaboration with C. Mueller (ETH), we prepared a dataset consisting of two classes of test data. The first class consists of challenging height functions used in optimization (Himmelblau, Rastrigin, Mueller-Brown, Monkey saddles, and various cases based on Gaussian mixtures). The second class consists of height functions defined by polynomials in three variables. Such examples are especially interesting since their critical points can be certified used real solving tools (Gröbner basis coupled to the Rational Univariate Representation).

From the implementation standpoint, our algorithm was coded in CGAL style i.e. generic C++, to accommodate both molecular and Euclidean data. The results are almost perfect on all the aforementioned test-cases. For comparison purposes in the 3D setting, we also implemented the discrete Morse-Smale algorithm presented in the paper *Theory and algorithms for constructing discrete morse complexes from grayscale digital images*, by Robins, Wood and Sheppard, IEEE PAMI 2011. We believe that comparing our approach which is solely based on points, against an algorithm in the realm of Forman's Morse theory is indeed of interest.

### Task 3.a: Sampling and approximation of manifolds

We developed further the asymptotic approximation of both the developable and ruled patches of the contact surfaces in the work space of a planar robot [2]. We next want to obtain an optimal approximation in the Hausdorff sense of these surface patches. We consider the vertical distance, and obtain an optimal local triangulation of hyperbolic paraboloid surfaces w.r.t. to this distance. We want to use this local triangulation, as a building block of the approximation of negatively curved surface patches in general.

As previously our focus has been on asymptotic approximations of manifolds (in particular surfaces), such as Fejes Tóth [8], Schneider [13] and Gruber [10, 11] approximate convex (hyper-) surfaces. In our effort towards a generalization of their results to arbitrary manifolds we have shown that there is no intrinsic geometric measure for the complexity of an approximation (see CGL-TR-42). This is heuristically clear because the rigidity of a manifold disappears if the codimension of the embedding is sufficiently high, as was noted by Nash in his seminal paper [12]. In fact a compact $n$-manifold has a isometric embedding in any small volume of Euclidean $(n/2)(3n + 11)$-space. So roughly speaking, one can squash a manifold in a small volume without affecting the curvatures but this would lead to 'wrinkles'. This result justifies the approach used previous work of in our group on the generalization of result by Clarkson [7], see the Master's thesis of David de Laat submitted in year 1.

Furthermore, we are working on a general result regarding the complexity of the approximation of a surface by means of piecewise quadratic patches. Such quadratic surfaces are characterized by having a constant Pick-invariant, an intrinsic property of surfaces in affine differential geometry. The intricate expressions occurring in our computations so far prevent a generalization of our earlier complexity result for piecewise linear approximations discussed in last year's contribution CGL-TR-10.

### Task 3.b: Upsampling using Delaunay refinement

No work proposed in Period 2.

## Milestones

### MS 4: Decide on further developments based on preliminary results on real data

*Geometric inference:* the distance to measures framework introduced during Period 1 have shown to be very useful to address both geometric inference and statistical problems (see WP1 report for Period 1). However, also the computation of the distance to an empirical measure only relies on nearest neighbors computations, the efficient computation or approximation of its topological persistence remains a difficulty. We intend to work on the design of filtered complexes and approximation of the distance to measure functions in order to efficiently approximate its topological features.

*Clustering:* our preliminary implementation and experiments for the algorithm to cluster probabilistic data show that, so far, the algorithm can only deal with instances of medium size. To improve efficiency, we will also introduce heuristical approaches for the subprocedures it uses. On another hand, our persistence based clustering algorithm (see WP1 report for Period 1) can be very efficiently implemented and has proven to give very good results on large data sets. As it comes with theoretical guarantees for the stability of the number of clusters we intend to further explore the question of the clusters themselves.

### MS 5: Assess geometric criteria to identify inherent scales

Despite the promising statistical results obtained using model selection for simplicial complexes, the question of selecting a single relevant geometric scale remains widely open. In particular, understanding the connection between the relevant statistical scales (i.e. the ones obtained through a statistical criterion) and the geometric ones is one of the main difficulty that need to be addressed. On another hand, data often carry relevant geometric structures at different scales. We thus have decided to focus our efforts to an approach based upon topological persistence that allows to overcome the problem of selecting a single relevant scale. The stability results obtained in this direction provide efficient methods for geometric multiscale data analysis and open the door for new statistical criteria to select multiple relevant scales. We thus intend to design persitence based criteria during Period 3.

# Bibliography


[1] Mridul Aanjaneya, Frederic Chazal, Daniel Chen, Marc Glisse, Leonidas Guibas, and Dmitriy Morozov. Metric graph reconstruction from noisy data. *to appear in Intern. Journal on Computational Geometry and Applications*, 2012.

[2] Dror Atariah and Günter Rote. Configuration space visualization. In *Proceedings of the 2012 symposuim on Computational Geometry*, SoCG '12, pages 415–416, New York, NY, USA, 2012. ACM.

[3] Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry A. Wasserman. Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72, 2012.

[4] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *J. Foundations of Computational Mathematics*, 11.

[5] F. Chazal, V. de Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. *arXiv:1207.3674*, 2012.

[6] F. Chazal, V. de Silva, and S. Oudot. Persistence stability for geometric complexes. *arXiv:1207.3885*, 2012.

[7] K.L. Clarkson. Building triangulations using $\epsilon$-nets. *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 2006.

[8] L. Fejes Tóth. *Lagerungen in der Ebene, auf der Kugel und im Raum.* Berlin, Göttingen, Heidelberg: Springer, 1953.

[9] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *To appear in Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013.

[10] P.M. Gruber. Assymptotic estimates for best and stepwise approximation of convex bodies I. *Forum Mathematicum*, 5:281–297, 1993.

[11] P.M. Gruber. Assymptotic estimates for best and stepwise approximation of convex bodies II. *Forum Mathematicum*, 5:521–538, 1993.

[12] John Nash. The imbedding problem for riemannian manifolds. *The Annals of Mathematics*, 63(1):pp. 20–63, 1956.

[13] R. Schneider. Zur optimalen approximation konvexer hyperflächen durch polyeder. *Mathematische Annalen*, 256:289–301, 1981.

[14] Donald R. Sheehy. Linear-size approximations to the vietoris-rips filtration. In *SOCG: Proceedings of the 28th ACM Symposium on Computational Geometry*, 2012.

[15] Donald R. Sheehy. A multicover nerve for geometric inference. In *CCCG: Canadian Conference in Computational Geometry*, 2012.