Project number IST-25582

CGL

Computational Geometric Learning

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

STREP

Information Society Technologies

Period covered:	November 1, 2010–October 31, 2011
Date of preparation:	October 30, 2011
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 Jena, Germany

In the following we describe the work done within Work Package 1 within the first 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

During the last decade, the study of the mathematical properties of distance functions to compact sets has provided a powerful framework to infer topological and geometric properties of unknown shapes from data sampled in a small neighborhood of them. Building on this framework and considering data as empirical probability measures rather than just compact sets, we have introduced and developed new distance-based tools for geometric inference [6]. Indeed we have defined a way to associate to any probability distribution a function that shares fundamental mathematical properties with usual distance functions and whose sublevel sets carry geometric information on the distribution. We thus have obtained results allowing to infer the geometric information from data corrupted by noise and outliers. Our results apply in any dimension and the computation of the distance-like functions we have introduced boils down to the computation of nearest neighbors in the data set. Our results have already been used and applied in different settings:

- in statistics to design new consistent density estimates [1]. These estimates have (up to reparametrization) the same level sets as our distance-like functions and we have proven that they converge to the actual density from which the data are sampled. As a consequence we

have obtained the first density estimates that come with statistical convergence and geometric inference guarantees.

- in GIS to extract information from GPS data [5]. We have introduced a method that consists in embedding GPS traces data (that are represented as sequences of points in the plane) into a higher dimensional spaces and then to use the gradient of the distance-like functions to "smooth" the data and remove noise and measurements errors.

- combining our approach to deconvolution tools we have shown that the geometric information still can be inferred from data corrupted with a large amount of noise when the nature of the noise is known [3];

- using the framework of distance-like functions we have also initiated an on-going work on the inference of filamentary structures in astronomic data (see Task 3.a in Work Package 3).

Task 1.b: Clustering

Regarding clustering, the goal was both to improve some classical clustering algorithms for probabilistic input and to design new clustering approaches based upon topological approaches.

We have developed and analyzed a core-set construction for the k-median clustering problem with probabilistic data [8]. The construction can be used to obtain a $(1+\epsilon)$-approximation algorithm as well as a streaming algorithm for this problem.

We proved that there is a linear oblivious embedding for d-dimensional subspaces of l_1 into $O(d \log d)$-dimensional l_1 with distortion $O(d \log d)$ [9]. We showed that this embedding can be used to develop improved streaming and approximation algorithms for l_1 regression. The technique may also be helpful to obtain approximation algorithms for related clustering problems.

In a different setting we have used the theory of topological persistence introduced a decade ago by H. Edelsbrunner et al to design a new clustering algorithm that comes with theoretical guarantees [7]. Our algorithm can be seen as a variant of the classical "hill-climbing" algorithms that intend to cluster the data according to the basins of attraction of the local maxima of the density from which they have been generated. The use of topological persistence provides new robust tools to help the user to robustly determine the relevant number of clusters in the data and to get rid of non relevant clusters due to noise. In practice our algorithm relies on a basic variant of the union-find data structure. As a consequence it can be easily and efficiently implemented to process large data sets.

Task 2.a: Scale selection

Geometric inference is inherently multiscale (the topological structure of data usually depends on the scale at which these data are considered). We have adopted statistical approaches to try to automatically select relevant scales at which the data should be considered. In this direction we have continued and achieved a work initiated before the beginning of the CG-Learning project. We have designed a penalized criterion to select a relevant simplicial complex among a family of complexes approximating the considered data at different scales [4]. The validity of this criterion is assessed in a statistical setting using the theory of model selection. We expect to explore applications of this method during the next period.

Task 2.b: Morse-Smale and flow complexes

No work proposed in Period 1.

Task 3.a: Sampling and approximation of manifolds

In order to focus our effort we first considered the class of *ruled surfaces* which arise in the context of contact surfaces of a polygonal planar robot and a polygonal obstacle. Here we strive to obtain an approximation which is optimal in the *Hausdorff distance* sense. Furthermore, we try to exploit the unique geometrical properties of the surfaces under consideration.

So far, we obtained asymptotic results on the approximation error for the helices which generate the contact surfaces. We expect to get better approximations by looking at the surfaces directly.

Progress has been made on polyhedral approximations of submanifolds in Euclidean Spaces. Our main objective was to find an optimal triangulation of a surface, that is using as few points as possible to achieve good accuracy in the sense of the Hausdorff distance. We restricted ourselves to the asymptotic setting, where the number of vertices is large. See techreport [10]. In the continuation of works by Schneider and Gruber (convex case) and Clarkson, our main focus has been on the asymptotic complexity of such approximations with vertices on the manifold. One of the major problems one faces when considering non-convex hypersurfaces is that the edge length is no longer guaranteed to decrease if the Hausdorff distance between the surface and its approximating mesh decreases. We have provided strong evidence indicating that for an arbitrary surface the maximal edge-length of an optimal approximating mesh goes to zero with the Hausdorff distance. This enabled us to establish a lower bound on the complexity of an optimal approximating mesh. We also obtained a result concerning to the complexity of arbitrary meshes for convex hypersurfaces. Furthermore, we considered examples of non-convergence of the area of approximating meshes of surfaces, in a sense a generalization of the Schwarz lantern.

Task 3.b: Upsampling using Delaunay refinement

We have proposed an algorithm that can mesh any smooth submanifold of Euclidean space. The algorithm is a variant of Delaunay refinement. It constructs the tangential complex and therefore avoids constructing any subdivision of the ambient space, which leads to a complexity that only depends linearly on the ambient dimension [2].

Milestones

MS1: Noise removal vs. effect minimization

In general, noise removal or outliers characterization is an ill-defined problem in data analysis and geometric inference does not escape this issue. We have shown that the framework of distance-like functions associated to probability measures allows to minimize the effect of outliers without having to exhibit them explicitly. However we have also shown that when the nature of the noise is known, it can be removed from the data. As in many cases the exact nature of the noise in data is unknown we have decided to concentrate our effort on methods that minimize the effect of the noise and outliers rather than trying to remove them.

MS2: Geometric criteria for scale selection

Statistical model selection provides an interesting framework for automatic scale selection but there remain many open questions. We have decided to pursue these and to consider other statistical approaches as well.

MS3: Choice of criteria for implementing constraint surfaces

For the three-dimensional configuration space for rigid robots in the plane, we will capitalize on the structure of the ruled surface and consider special methods that may produce long skinny triangles along the surface, but which are refined near the seams (edges). With the view of intersecting the mesh with arbitrary planes, as required by the hybrid representation of configuration spaces (Manifold Sampling, see Workpackage 3), this is preferable to a uniform mesh.

For general surfaces (for example in 6 dimensions), we intend to use the method of Clarkson [2006] as planned originally, to get an approximating mesh with samples on the surface, and then offset this approximation to both sides to get one-sided (conservative) approximations. The parametric form in which the contact surfaces are given lends itself well to obtaining the curvature estimates that are needed in Clarkson's method.

Bibliography

[1] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodrguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011.

[2] J.-D. Boissonnat and A. Ghosh. Triangulating smooth submanifolds with light scaffolding. *Mathematics in Computer Science*, 4(4):431–462, 2011.

[3] C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423, 2011.

[4] C. Caillerie and B. Michel. Model selection for simplicial approximation. *Accepted for publication in Foundations of Computational Mathematics*, 2011. Preliminary version available as INRIA research report.

[5] F. Chazal, D. Chen, L. Guibas, X. Jiang, and C. Sommer. Data-driven trajectory smoothing. In *Proceedings of the 19th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, 2011.

[6] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *J. Foundations of Computational Mathematics*, 2011. to appear (DOI) 10.1007/s10208-011-9098-0.

[7] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. In *Proc. 27th Annu. ACM Sympos. on Comput. Geom.*, pages 97–106, June 2011.

[8] C. Lammersen, M. Schmidt, and C. Sohler. Probabilistic k-median clustering in data streams. *CGL Technical Report*, 3, 2011.

[9] C. Sohler and D. P. Woodruff. Subspace embeddings for the l_1-norm with applications. In *Proceedings of the 43rd ACM Symposium on Theory of Computing*, pages 755–764, 2011.

[10] G. Vegter and M.H.M.J. Wintraecken. On the complexity of polyhedral approximations of submanifolds of euclidean spaces. Technical Report CGL-TR-10, University of Groningen, 2011.