Data Analysis using Computational Topology and Geometric Statistics

Peter Bubenik (Cleveland State University),
Gunnar Carlsson (Stanford University),
Peter T. Kim (University of Guelph)

Mar 8 – Mar 13, 2009

1 Overview of the Field

Mathematical scientists of diverse backgrounds are being asked to apply the techniques of their specialty to data which is greater in both size and complexity than that which has been studied previously. Large, high-dimensional data sets, for which traditional linear methods are inadequate, pose challenges in representation, visualization, interpretation and analysis. A common finding is that these massive data sets require the development of new theory and that these advances are dependent on increasing technical sophistication. Two such data-analytic techniques that have recently developed independently of each other have come to the fore, namely, Geometric Statistics and Computational Topology. Although the former uses geometric arguments, while the latter uses algebraic-topological arguments, and hence they appear disparate, there is substantial commonality and overlap just as in the more traditional overlap between geometry and topology. Thus the purpose of this workshop is to bring together these two research directions and explore their overlap, particularly in the service of statistical data analysis.

A standard paradigm assumes that the data comes from some underlying geometric structure, such as a curved submanifold or a singular algebraic variety. The observed data is obtained as a random sample from this space, and the objective is to statistically recover features of the underlying space and/or the distribution that generated the sample.

In Geometric Statistics one uses the underlying Riemannian structure to recover quantitative information concerning the probability distribution and/or functionals thereof. The idea is to extend statistical estimation techniques to functions over Riemannian manifolds, utilizing spectral methods adapted to the Riemannian structure.

One then considers the magnitude of the statistical accuracy of these estimators. Considerable progress has been achieved in terms of optimal estimation in the minimax sense. These ideas have far reaching implications in the analysis of high-dimensional data such as, for example, in astronomy, biomechanics, medical imaging, microwave engineering and texture analysis.

In Computational Topology, one attempts to recover more qualitative global features of the underlying data instead, such as connectedness, or the number of holes, or the existence of obstructions to certain constructions, based upon the random sample. In other words, one hopes to recover the underlying topology. An advantage of topology is that it is stable under deformations and thus insensitive to errors introduced in the sampling.

A combinatorial construction such as the alpha complex or the Čech complex converts the discrete data into an object for which it is possible to compute the topology. However, it is quickly apparent that such a construction and its calculated topology depend on the scale at which one considers the data. A multiscale solution to this problem is the technique of persistent homology. It quantifies the persistence of topological features as the scale changes. Persistent homology is useful for visualization, feature detection and object recognition. It has been successfully applied to analyze natural images, neurological data, gene-chip data, protein binding and sensor networks.

Although Geometric Statistics and Computational Topology have a disparate appearance and seem to have different objectives, it has recently been noticed that they share a commonality through statistical sampling. In particular it has been noticed that the metric distance of persistent homology in Computational Topology, is intimately related to the sup-norm metric between the underlying density that generates a random sample on a Riemannian manifold, and its statistical estimator. Consequently, the qualitative and quantitative data analyses are intimately linked, which is not surprising because of the close connection between geometry and topology traditionally.

2 Recent Developments and Open Problems

The use of geometric and topological methods for statistical data analysis is currently being pursued in the three allied fields of computer science, mathematics and statistics. Although each field has their own particular approach and questions of interest, the amount of similarity is striking and this workshop was able to synthesize all three fields together. The open problems that were considered was the development of computational and statistical algorithms and methods using aspects of geometry and topology when data over the geometric object was only available.

We can summarize the type of investigations as it pertains to the aforementioned three fields. A more detailed description is provided in the following section:

- In computer science the pursuit naturally focused on efficient algorithms and visualization. Some specific items discussed included algorithms for the discrete approximation of the Laplacian, algorithms for approximating cut-locus, data reduction techniques, and recovery from noisy data;

- In mathematics the interest focused on certain constructions. Here such topics included zigzag persistence, Hodge theory, and recovering the topology over a random field;

- In statistics parameter estimation was the main interest and topics included bootstrapping and MCMC on manifolds, geodesic PCA, asymptotic minimaxity, conditional independence, statistical multiscale analysis and analysis over the Euclidean motion group.

Additionally, some physical applications were also discussed such as brain mapping, network analysis and biomechanics of osteoarthritis.

3 Presentation Highlights (in alphabetical order)

Dominique Attali (CNRS, Grenoble)
Persistence-sensitive simplification of functions on surfaces in linear time.
Let f be a real-valued function defined on a triangulated surface S. The persistence diagram of f encodes the homological variations in the sequence of sublevel sets $S_t = f^{-1}(-\infty, t]$. A point (x, y) in the persistence diagram of f corresponds to a homological class which appears in S_x and disappears in S_y. The distance $y - x$ of the point (x, y) to the diagonal represents the importance of the associated homological class: the further away a point is from the diagonal, the more important the associated feature. An ε-simplification of f is a map g on S whose persistence diagram consists only of those points in the diagram of f that are more than ε away from the diagonal. The speaker gave an algorithm for constructing an ε-simplification of f which is also ε-close to f. This was a joint work with M. Glisse, S. Hornus, F. Lazarus and D. Morozov.

Peter Bubenik (Cleveland State University)
Persistent homology and nonparametric regression.

The talk focused on estimating the persistent homology of sublevel sets of a function on a compact Riemannian manifold, from a finite noisy sample. The Stability Theorem of Cohen-Steiner, Edelsbrunner and Harer bounds the distance between the persistent homologies of the sublevel sets of two functions by the supremum norm of the difference between the two functions. Using this result, the above topological problem was translated to the statistical nonparametric regression problem on a compact manifold under the sup-norm loss. The main result was a calculation the sharp asymptotic minimax bound. Furthermore, the construction of the estimator in the proof is well-suited to calculations of the persistent homology of its sublevel sets. These techniques were applied to brain image data. Initial results indicated the possibility of distinguishing autistic and control subjects by the topology of their brains. This was joint work with Gunnar Carlsson, Moo Chung, Peter Kim, and Zhiming Luo.

Gunnar Carlsson (Stanford University)
Generalized Persistence, Noise, and Statistical Significance
Persistent homology has been shown to be a useful way to detect qualitative structure in various kinds of data sets. The speaker showed that a generalized form of persistence, called “zig-zag persistence”, can be useful both in removing noise in certain geometric problems as well as in understanding statistical significance of qualitative geometric invariants. This was joint work with V. de Silva and D. Morozov.

Fred Chazal (INRIA)
Geometric inference for probability distributions
Data often comes in the form of a point cloud sampled from an unknown compact subset of Euclidean space. The general goal of geometric inference is then to recover geometric and topological features (Betti numbers, curvatures,. . .) of this subset from the approximating point cloud data. In recent years, it has appeared that the study of distance functions allows one to address many of these questions successfully. However, one of the main limitations of this framework is that it does not cope well with outliers nor with background noise. The speaker showed how to extend the framework of distance functions to overcome this problem. Replacing compact subsets by measures, he introduced a notion of distance function to a probability distribution in $\mathbb{R}^n$. These functions share many properties with classical distance functions, which makes them suitable for inference purposes. In particular, by considering appropriate level sets of these distance functions, it is possible to associate in a robust way topological and geometric features to a probability measure. This was joint work with David Cohen-Steiner.

Moo Chung (University of Wisconsin-Madison)
Eigenfunctions of Laplace-Beltrami operator in cortical manifolds
In quantifying cortical and subcortical anatomy of the human brain, various differential geometric methods have been proposed. Many such successful methods are inherently implicit and without explicit parametric forms. Although there are a few parametric approaches such as spherical harmonic descriptors, their application has been limited to simple subcortical structures. The reason for the lack of more explicit parametric approaches is that it is difficult to construct an orthonormal basis for an arbitrary cortical manifold. The speaker proposed to use the eigenfunctions of the Laplace-Beltrami operator, which are computed numerically using the cotan formula. The eigenfunctions are then used in setting up a regression in the cortical manifold. In the heat kernel smoothing framework, smoothing is done by expanding the heat kernel using the eigenfunctions. The eigenfunction approach offers far more flexibility in setting up a statistical model than implicit approaches.

Vin de Silva (Pomona)
Zigzag persistence
Zigzag persistence is a new methodology for studying persistence of topological features across a family of spaces or point-cloud data sets. Building on classical results about quiver representations, zigzag persistence generalises the highly successful theory of persistent homology and addresses several situations which are not covered by that theory. The speaker presented theoretical and algorithmic foundations with a view towards applications in topological statistics. As an important example, he discussed a particular zigzag sequence derived from the level sets of a real-valued function on a topological space. A powerful structure theorem, called the Pyramid Theorem, establishes a connection between this ”levelset zigzag persistence” and the

extended persistence of Cohen-Steiner, Edelsbrunner and Harer. This theorem resolves an open question concerning the symmetry of extended persistence. Moreover, the interval persistence of Dey and Wenger can be understood in this context; in some sense it carries three-quarters of the information produced by the other two theories. This was joint work with Gunnar Carlsson and Dmitriy Morozov.

Tamal Dey (Ohio State University)
Topology by approximating cut locus from point data
A cut locus of a point p in a compact Riemannian manifold M is defined as the set of points where *minimizing* geodesics issued from p stop being minimizing. It is known that a cut locus contains most of the topological information of M. One can try to utilize this property of the cut loci to decipher the topology of M from a point sample. Recently it has been shown that Rips complexes can be built from a point sample P of M systematically to compute the Betti numbers, the rank of the homology groups of M. Rips complexes can be computed easily. However, the sizes of the Rips complexes tend to be large. Since the dimension of a cut locus is lower than that of the manifold M, a sub-sample of P approximating the cut locus is usually much smaller in size and hence admits a relatively smaller Rips complex. The speaker explored the above approach for point data sampled from surfaces embedded in any high dimensional Euclidean space. He presented an algorithm that computes a sub-sample P' of a sample P of a 2-manifold where P' approximates a cut locus. Empirical results show that the first Betti number of M can be computed from the Rips complexes built on these sub-samples.

Leo Guibas (Stanford)
Analysis of Scalar Fields over Point Cloud Data
Given a real-valued function f defined over some metric space X, is it possible to recover some structural information about f from the sole information of its values at a finite subset L of sample points, whose pairwise distances in X are given? The speaker provided a positive answer to this question. More precisely, taking advantage of recent advances on the front of stability for persistence diagrams, he introduced a novel algebraic construction, based on a pair of nested families of simplicial complexes built on top of the point cloud L, from which the persistence diagram of f can be faithfully approximated. He then derived from this construction a series of algorithms for the analysis of scalar fields from point cloud data. These algorithms are simple and easy to implement, have reasonable complexities, and come with theoretical guarantees. This was joint work with F. Chazal, S. Y. Oudot, and P. Skraba.

Susan Holmes (Stanford)
How to sample from a manifold: Applications to validation of Computational Topology and its algorithms
The speaker surveyed the classical methods of parametric bootstrapping and MCMC for generating samples from non uniform distributions. Then she presented work on how to draw samples from a manifold and show how this can be used to compute confidence statements for results from various outputs from computational topology algorithms such as JPlex. This was joint work with Persi Diaconis and Mehrdad Shahshahani.

Stephan Huckeman (Goettingen)
Intrinsic Statistics on Riemannian Manifolds
One goal in image analysis consists in describing statistical distributions of characteristic patterns, e.g. shapes of random physical objects. Typically such shapes live on non-Euclidean manifolds, possibly with unbound curvature at singularities (e.g. Kendall's 3D shape space). While over the last decades statisticians have used Euclidean approximations to these manifolds thus making tools of classical multivariate analysis available for "sufficiently concentrated data", this talk aimed at intrinsic generalizations of PCA and MANOVA thus broadening the scope of statistical image analysis.

Matt Kahle (Stanford University)
Moduli spaces of hard disks in a box
The speaker discussed a family of moduli spaces which generalize classical configuration spaces for points in the plane. However, the methods used for computing homology of configuration spaces are not easily applicable to these spaces, and even the number of components seems to be a fairly subtle question. So computational / applied methods were used to better understand this pure math problem. A combination of

techniques, including simulated annealing and the nudged elastic band method, were used to compute the most basic topological features of these spaces. There was also a brief discussion of the statistical physics setting that motivates the problem, suggested by Persi Diaconis. This was ongoing joint work with Gunnar Carlsson and Jackson Gorham.

Andre Lieutier (Dassault Systemes)
A stable notion of curvature on point clouds
The speaker addressed the problem of curvature estimation from sampled compact sets. The main contribution was a stability result: the gaussian, mean or anisotropic curvature measures of the offset of a compact set K with positive μ-reach can be estimated by the same curvature measures of the offset of a compact set K' close to K in the Hausdorff sense. He showed how these curvature measures can be computed for finite unions of balls. The curvature measures of the offset of a compact set with positive μ-reach can thus be approximated by the curvature measures of the offset of a point-cloud sample. These results can also be interpreted as a framework for an effective and robust notion of curvature. This was joint work with Frederic Chazal, David Cohen-Steiner and Boris Thibert.

Zhiming Luo (University of Guelph)
Asymptotic minimax regression estimate under super-norm loss on Riemannian manifold
Relating to Peter Bubenik's talk "Persistent homology and nonparametric regression", the speaker gave more details on the minimax nonparametric regression estimator and the exact constant of the sharp asymptotic minimax bound on a compact Riemannian manifold. This was joint work with Peter Bubenik, Gunnar Carlsson, Moo Chung, and Peter Kim.

Facundo Memoli (Stanford)
A Metric Geometry approach to Object Matching
The problem of object matching under invariances can be studied using certain tools from Metric Geometry. The main idea is to regard objects as metric spaces (or measure metric spaces). The type of invariance one wishes to have in the matching is encoded in the choice of the metrics with which one endows the objects. The standard example is matching objects in Euclidean space under rigid isometries: in this situation one would endow the objects with the Euclidean metric. More general scenarios are possible in which the desired invariance cannot be reflected by the preservation of an ambient space metric. Several ideas due to M. Gromov are useful for approaching this problem. The speaker discussed different adaptations of these, and in particular he constructed an L^p version of the Gromov-Hausdorff distance using mass transportation ideas.

Yuriy Mileyko (Duke University)
Defining hierarchical order within reticular networks
While the Strahler Stream Order is a standard method for computing the hierarchical order within non-reticular networks, it cannot handle networks with loops. The speaker presented a new algorithm which can perform such a task for planar networks. This algorithm is based on ideas from persistent homology and may be regarded as a generalization of the Strahler Stream Order. From a topological point of view, the latter method defines a filtration of a network (based on tributaries) and updates the order of the edges at critical events, that is, when two connected components merge. Such an event can be regarded as a change in 0-dimensional homology. Therefore, he defined critical events for networks with loops as changes in 1-dimensional homology. Taking advantage of the planarity of a network, one can trace a sequence of such critical events and update the order of network edges. This work was motivated by the problem of analyzing the structure of leaf networks, and the speaker presented a few preliminary results of such an analysis. He also discussed possible generalizations of the new method to arbitrary networks.

Konstantin Mischaikow (Rutgers University)
Topology Guided Sampling of Nonhomogeneous Random Fields
Topological measurements are increasingly being accepted as an important tool for quantifying complex structures. In many applications these structures can be expressed as nodal domains of real-valued functions and are obtained only through experimental observation or numerical simulations. In both cases, the data on which the topological measurements are based are derived via some form of finite sampling or discretization.

The speaker presented a probabilistic approach to quantifying the number of components of generalized nodal domains of non-homogeneous random fields in one space dimension via finite discretizations, i.e., he considered excursion sets of a random field relative to a non-constant deterministic threshold function. He gave explicit probabilistic a-priori bounds for the suitability of certain discretization sizes and also provided information for the choice of location of the sampling points in order to minimize the error probability. He illustrated the results for a variety of random fields, demonstrated how they can be used to sample the classical nodal domains of deterministic functions perturbed by additive noise, and discussed their relation to the density of zeros.

Sayan Mukherjee (Duke University)
Conditional Independence Models via Filtrations
The speaker presented a novel approach to infer conditional independence models or Markov structure of a multivariate distribution. Specifically, the objective is to place informative prior distributions over decomposable graphs and sample efficiently from the induced posterior distribution. The key idea is a parametrization of decomposable hypergraphs using the geometry of points in $\mathbb{R}^m$. This allows for the specification of informative priors on decomposable graphs by priors on a finite set of points. The constructions used have been well studied in the fields of computational topology and random geometric graphs. The framework underlying this idea was developed and its efficacy was illustrated using simulations.

Axel Munk (Goettingen)
Statistical Multiscale Analysis - From Jump detection to Image Analysis
The speaker discussed how to use statistical multiscale analysis (SMA) techniques in order extract jumps from noisy signals in various signal detection problems. This was applied to reconstruct the open states in ion channel experiments for biomembranes. In the second part SMA was extended to image analysis, i.e. to 2D and 3D. The resulting method is locally adaptive, i.e. it automatically adjusts locally any regularisation method to locally varying features, such as edges. This was illustrated with examples from biophotonic imaging.

Vic Patrangenaru (Florida State University)
Asymptotic Statistics and Nonparametric Bootstrap on Manifolds and Applications
Asymptotic statistical analysis and nonparametric bootstrap on smooth geometric objects, or manifolds, is an exciting and challenging field of research, extending multivariate limit theorems to the nonlinear case, where statistical theory and differential geometry are inextricably intertwined, and implementation requires innovative algorithms and high speed computation. This presentation dealt with recent developments in this young area of nonparametric statistics, which must also resolve associated geometric issues and problems of implementation. Asymptotic statistics on manifolds have a wide range of applications in many areas of science including geology, meteorology, biology, medical imaging, bioinformatics and machine vision. This was joint work with R. N. Bhattacharya, F. H. Ruymgaart and other collaborators.

Michael Pierrynowski (McMaster University)
Differential geometry reveals differences in the knee motion of elders with osteoarthritis
Knee motion, force and moment have been used by biomechanists to identify elders with and without knee osteoarthritis (OA). The knee adduction moment has received the most attention since it is associated with the severity and prognosis of OA which then informs clinicians to prescribe effective intervention. However, measuring the knee adduction moment clinically is problematic since it requires synchronized kinematic data acquisition and ground reaction force measures. For potential clinical use the speaker proposed a differential geometry analysis of the easier measured knee kinematics [SE(3)] that shows promise to detect the presence or absence of mild to moderate knee OA. This technique sums over repetitive gait cycles the curvatures and torsions from the translation component of SE(3) which are then geometrically interpreted using Frechel's Theorem. In a similar vein, he examined the length of the paths transcribed on a sphere (S^2) by the three columns (orthonormal vectors) of the SO(3) orientation component. He reported that during repetitive normal overground gait, the sum of the curvatures and the path length of the third SO(3) vector are smaller in 52 elders with knee osteoarthritis compared to 47 elders with healthy knees. He discussed this finding in relation to OA knees having decreased non-linear motion paths and less tibia rotation during gait. This was joint work with Peter T. Kim.

Louis-Paul Rivest (Université Laval)
Some statistical models for SE(3) data
This presentation began by reviewing the occurrence, in the biomechanical literature, of data sets whose elements belong to SE(3), the 6-dimensional Lie group of 3D rigid body displacements. The construction of some probability models on SE(3) using distance measures was presented. These models were used to describe the dispersion of an observed SE(3) displacement around its "true value". They were used to construct loss functions for the estimation of the parameters of a statistical model for SE(3) data. The SE(3) model used to estimate the directions of the two rotation axes of the ankle was then be presented. Some statistical challenges associated with the estimation of the parameters of this model were reviewed, with some of the solutions that have been put forward. Statistical analyses carried out with the R-package Kinematics for the statistical modeling of SE(3) data were be used to illustrate the theory.

Stephen Smale (Toyota Technological Institute at Chicago)
Hodge Theory
The speaker discussed results on extensions of Hodge theory to metric spaces and the relations to the subject "Topology, Geometry and Data". This was joint work with Nat Smale.

Mikael Vejdemo-Johansson (Stanford University)
Persistent Cohomology and Circular Coordinates
An inherent assumption in algorithms for linear or non-linear dimension reduction (NLDR) is that the data will be representable faithfully and efficiently using real-valued coordinates. However, there are examples that challenge this assumption: the circle, for instance, being inherently one-dimensional, but using two real coordinates for a faithful representation. The speaker presented a strategy for constructing circle-valued functions on a statistical data set. He developed a machinery of persistent cohomology to identify candidates for significant circle-structures in the data, and used harmonic smoothing and integration to obtain circle-valued coordinate functions from representative cocycles of the cohomology classes recovered. He suggested that the enriched class of either real- or circle-valued coordinate functions permits a precise NLDR analysis of a broader range of realistic data sets.

Yusu Wang (Ohio State University)
Approximating Laplace-Beltrami Operator, Integrals and Gradients in Non-statistical Discrete Settings
The Laplace-Beltrami operator of a given manifold (e.g, a surface) is a fundamental object encoding the intrinsic geometry of the underlying manifold. It has many properties useful for practical applications from areas such as graphics and machine learning. For example, its relation to the heat diffusion makes it a primary tool for surface smoothing in graphics. However, many a time, the underlying manifold is only accessible through a discrete approximation, either as a mesh or simply as a set of points. The important question is then how to approximate the Laplace operator and other geometric invariants from such discrete setting. Previously, much work has been done on approximating Laplace operator from points sampled from some probabilistic distribution. The speaker described her recent results on approximating the Laplace operator from either piecewise linear manifolds (e.g, meshes) or simply general point cloud data. She then gave several applications of the constructed discrete Laplace operator, including estimating the gradient, critical points, and the integral of an input function from point cloud data.

4 Scientific Progress Made

The lectures were organized so that there was approximately equal representation from each of the fields of computer science, mathematics and statistics. Throughout the week, each lecture was well attended with much discussion and enthusiasm displayed by the audience. Questions and inquiries were made from scientists within each field, but also from the other fields as well. The scientific progress that was most prominent therefore can be summarized in terms of the cross-fertilization between the above three fields, and the attempt to bridge the "language gap" between the three fields. The benefits of this synergy was carried on much beyond the lectures, and what was particularly fascinating was that in some fields, what they had been struggling with, was something fairly well known in others.

As mentioned in the introduction, an important goal was broadening the horizons of statistical data analysis. It terms of this it is evident that the statisticians have considerable advantages. What is missing from statisticians is the plethora of geometrical and topological techniques as neither is typically carried out in graduate statistical training. In particular, development of statistical techniques usually take place over Euclidean space. On the other hand, computer scientists and mathematicians have the geometric and topological training, as well as experience in development of algorithms. Nevertheless, statistical techniques needed in non-Euclidean settings are limited consequently, statistical expertise for these types of structures are not readily available as compared to that available for Euclidean space. Consequently, significant scientific progress was made in communicating ideas of each others fields in the pursuit of using geometric and topological methods for statistical data analysis.

5 Outcome of the Meeting

Throughout the meeting, the organizers informally and continually canvassed the participants as to their thoughts on the progress of the workshop thus far. A common approval by participants was the sentiment. On the final day of the workshop, an informal discussion was held to discuss the outcome of the meeting as well as possible future like events. A very enthusiastic approval was relayed by all participants along with a very sincere desire going forward, to have more of such meetings either as a BIRS workshop, or other formats and venues. It was expressed that an equal representation from computer science, mathematics and statistics is the ideal mixture. Furthermore, in future meetings, in addition to the methodological advances made, a greater quantity of physical applications using geometric and topological statistical data analysis techniques was desired.

References

[1] D. Attali, H. Edelsbrunner, and Y. Mileyko, Weak witnesses for delaunay triangulations of submanifolds. In *SPM '07: Proceedings of the 2007 ACM symposium on Solid and physical modeling*, 143–150, ACM Press, New York, NY, USA, 2007.

[2] J.F. Angers and P.T. Kim, Multivariate Bayesian function estimation, *Ann Statist* **33** (2005), 2967-2999.

[3] R. Bhattacharya and V. Patrangenaru, Large sample theory of intrinsic and extrinsic sample means on manifolds – II, *Ann Statist* **33** (2005), 1225–1259.

[4] M. Belkin and P. Niyogi, Semi-Supervised Learning on Riemannian Manifolds, *Mach Learn* **56** (2004), 209-239.

[5] M. Belkin, J. Sun, and Y. Wang, Constructing laplace operator from point clouds in rd. In *SODA '09: Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, 1031–1040, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics, 2009.

[6] N. Bissantz, T. Hohage, A. Munk and F. Ruymgaart, Convergence rates of general regularization methods for statistical inverse problems and applications, *SIAM J Numerical Analysis* **45** (2007), 2610-2636.

[7] J.-D. Boissonnat, L.J. Guibas, and S.Y. Oudot, Manifold reconstruction in arbitrary dimensions using witness complexes. In *SCG '07: Proceedings of the twenty-third annual symposium on Computational geometry*, 194–203, New York, NY, USA, ACM Press, 2007.

[8] P. Bubenik and P.T. Kim, A statistical approach to persistent homology, *Homology Homotopy and Applications* **9** (2007), 337-362.

[9] G. Carlsson, Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, **46** (2009), 255–308.

[10] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, On the local behavior of spaces of natural images, *Int. J. Comput. Vision*, **76** (2008), 1–12.

[11] F. Chazal and A. Lieutier, Weak feature size and persistent homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples. In *SCG '05: Proceedings of the twenty-first annual symposium on Computational geometry*, 255–262, New York, NY, USA, ACM Press, 2005.

[12] F. Chazal and A. Lieutier, Smooth manifold reconstruction from noisy and non-uniform approximation with guarantees, *Comput. Geom.*, **40** (2008), 156–170.

[13] M.K. Chung, S. Robbins, R.J. Davidson, A.L. Alexander, K.M. Dalton, and A.C. Evans, Cortical thickness analysis in autism with heat kernel smoothing, *NeuroImage* **25** (2005), 1256–1265.

[14] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, Stability of persistence diagrams. In *SCG '05: Proceedings of the twenty-first annual symposium on Computational geometry*, 263–271, New York: ACM Press, 2005.

[15] F. Cucker, and S. Smale, On the mathematical foundations of learning, *Bull Amer Math Soc* **39** (2002), 1–49.

[16] V. de Silva, and G. Carlsson, Topological estimation using witness complexes, *Eurographics symposium on point-based graphics*, 2004.

[17] V. de Silva, and R. Ghrist, Homological sensor networks, *Notic Amer Math Soc* **54** (2007), 10–17.

[18] V. de Silva, and P. Perry, Plex version 2.5, available online, 2005.
`http://math.stanford.edu/comptop/programs/plex`

[19] T.K. Dey. *Curve and surface reconstruction: algorithms with mathematical analysis*, *Cambridge Monographs on Applied and Computational Mathematics*, **23**, Cambridge University Press, Cambridge, 2007.

[20] H, Edelsbrunner, and J. Harer, Persistent homology—a survey. In *Surveys on discrete and computational geometry*, *Contemp. Math.*, **453**, 257–282. Amer. Math. Soc., Providence, RI, 2008.

[21] H. Edelsbrunner, D. Letscher, and A. Zomorodian, Topological persistence and simplification, *Discrete Comput. Geom.* **28** (2001), 511-533.

[22] H. Edelsbrunner, M-L. Dequent, Y. Mileyko, and O. Pourquie, Assessing periodicity in gene expression as measured by microarray data. Preprint.

[23] H. Hendriks, Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions, *Ann Statist* **18** (1990), 832–849.

[24] S. Holmes, Bootstrapping phylogenetic trees: theory and methods, *Statist. Sci.*, 18 (2003), 241–255.

[25] P.T. Kim, J.Y. Koo, Statistical inverse problems on manifolds, *J Fourier Anal Appl* **11** (2005), 639–653.

[26] P.T. Kim, J.Y. Koo, and Z. Luo, Weyl eigenvalue asymptotics and sharp adaptation on vector bundles, *J Multivariate Anal.* accepted, 2009.

[27] J. Klemelä, Asymptotic minimax risk for the white noise model on the sphere, *Scand J Statist* **26**, 465–473.

[28] J.Y. Koo, and P.T. Kim, Asymptotic minimax bounds for stochastic deconvolution over groups, *IEEE Transactions on Information Theory* **54** (2008), 289 – 298.

[29] A.P. Korostelev, and M. Nussbaum, The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli* **5** (1996), 1099–1118.

[30] K.V. Mardia, and P.E. Jupp. *Directional statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester, 2000.

[31] K. Mischaikow, and T. Wanner, Probabilistic validation of homology computations for nodal domains, *Ann Appl Probab*, **17** (2007), 980–1018.

[32] S. Mukherjee, and Q. Wu, Estimation of gradients and coordinate covariation in classification, *J Mach Learn Res*, **7** (2006), 2481–2514.

[33] P. Niyogi, S. Smale, and S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, *Discrete Comput. Geom.*, **39** (2008), 419–441.

[34] B. Pelletier, Kernel density estimation on Riemannian manifolds, *Stat Prob Letter* **73** (2005), 297–304.

[35] B. Pelletier, Non-parametric regression estimation on a closed Riemannian manifold, *J Nonparametric Statist* **18** (2006), 57–67.

[36] G. Singh, F. Memoli, and G. Carlsson, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, 91–100, Prague, Czech Republic, Eurographics Association, 2007.

[37] S. Smale, and D-X. Zhou, Shannon sampling and function reconstruction from point values, *Bull Amer Math Soc* **41** (2004), 279-305.

[38] A.J. Zomorodian, *Topology for computing*, *Cambridge Monographs on Applied and Computational Mathematics* **16**, Cambridge University Press, Cambridge, 2005.

[39] A. Zomorodian, and G. Carlsson, Computing persistent homology, *Discrete Comput Geom* **33** (2005), 249–274.