Mathematisches Forschungsinstitut Oberwolfach

Report No. 3/2018

DOI: 10.4171/OWR/2018/3

Statistics for Data with Geometric Structure

Organised by
Aasa Feragen, Copenhagen
Thomas Hotz, Ilmenau
Stephan Huckemann, Göttingen
Ezra Miller, Durham

21 January – 27 January 2018

Abstract. Statistics for data with geometric structure is an active and diverse topic of research. Applications include manifold spaces in directional data or symmetric positive definite matrices and some shape representations. But in some cases, more involved metric spaces like stratified spaces play a crucial role in different ways. On the one hand, phylogenetic trees are represented as points in a stratified data space, whereas branching trees, for example of veins, are data objects, whose stratified structure is of essential importance. For the latter case, one important tool is persistent homology, which is currently a very active area of research. As data sets become not only larger but also more complex, the need for theoretical and methodological progress in dealing with data on non-Euclidean spaces or data objects with nontrivial geometric structure is growing. A number of fundamental results have been achieved recently and the development of new methods for refined, more informative data representation is ongoing. Two complimentary approaches are pursued: on the one hand developing sophisticated new parameters to describe the data, like persistent homology, and on the other hand achieving simpler representations in terms of given parameters, like dimension reduction. Some foundational works in stochastic process theory on manifolds open the doors to this field and stochastic analysis on manifolds, thus enabling a well-founded treatment of non-Euclidean dynamic data. The results presented in the workshop by leading experts in the field are great accomplishments of collaboration between mathematicians from statistics, geometry and topology and the open problems which were discussed show the need for an expansion of this interdisciplinary effort, which could also tie in more closely with computer science.

Mathematics Subject Classification (2010): 62xx, 53xx, 60xx, 65xx, 14xx.

Introduction by the Organisers

The workshop *Statistics for Data with Geometric Structure*, organized by Aasa Feragen (Copenhagen), Thomas Hotz (Ilmenau), Stephan Huckemann (Göttingen) and Ezra Miller (Durham) had 48 participants from many countries around the world. In particular, 14 of the 17 participants in the mini-workshop *Asymptotic Statistics on Stratified Spaces* held in 2014 at the MFO took part in this workshop. The interdisciplinary nature of the subject matter was reflected in the very diverse mathematical backgrounds of the speakers.

In the past years, data with geometric structure play an increasingly important role in statistics and lead to a surge in the application of geometric and topological concepts in statistical data analysis. Two major classes of approaches are pursued in this field. The first approach seeks to represent geometric objects as points in a non-Euclidean data space, while the second approach seeks to extract the major features of the geometric object to achieve a refined representation, not necessarily in a non-Euclidean space.

In the first approach, the spaces need not even be manifolds, but can be stratified spaces, in which case means can have non-standard properties, called stickiness [2] and repulsiveness. These are especially relevant for phylogenetic tree spaces which are used in population genetics. Calculation of geodesics [4] and analogues to principal components [3] is very challenging in these spaces.

On the other hand, measures on spaces with positive curvature can exhibit lower rates of asymptotic convergence of the sample mean, called smeariness [5, 6]. Such spaces of positive curvature are the principal object of concern in directional statistics and many shape representations.

For many spaces, refined methods have been developed, for example for dimension reduction and also some generic asymptotic results were achieved, see e.g. [7, 8, 9]. Furthermore, many specific difficulties for various data representations have been described and partly solved [12, 13].

A very important field, which is currently emerging, is the theory of stochastic processes and stochastic analysis on manifolds. Recent important foundational work has been done by Sommer and Joshi [11], whose collaboration was fostered by the mini-workshop *Asymptotic Statistics on Stratified Spaces*. The development of new models, the underlying computational theory, as well as computational tools are a milestone towards an effective treatment of stochastic processes on manifolds.

The second approach to data with geometric structure seeks to extract the major features of the geometric object to achieve a refined representation. A major technique to this effect is persistent homology (for an introduction and historical overview, see [1]), which is increasingly used in image and shape analysis. In this class of methods, scale-space-like transformations are used to represent complicated geometric objects in terms of topological properties. For example, separation of clusters and sizes of holes in a data set can be quantified in terms of persistence diagrams.

For every data set, the construction scheme of the persistence diagram must be reconsidered. In many applications, level sets of (possibly multivariate) functions

are considered; in some cases, objects are sliced in different angles to create a whole ensemble of persistence diagrams [14] and also several independent parameters, leading to higher dimensional persistence diagrams are considered (see the contribution by E. Miller).

Furthermore, the parameters of interest to extract from the persistence diagram must be determined for every application specifically. This can range from a reduction to simple scalar summary statistics, over curves [15] to sophisticated analysis applying tropical geometry [16].

The workshop provided an overview over the very diverse subject of statistics for data with geometric structure and a number of different ways to approach the subject. It initiated lively discussions concerning several topics, which were further discussed in five focus groups.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Ruriko Yoshida in the "Simons Visiting Professors" program at the MFO.

References

[1] H. Edelsbrunner and J. Harer, *Persistent Homology – a Survey*, in: Surveys on Discrete and Computational Geometry: Twenty Years Later (2008).

[2] T. Hotz, S. Huckemann, H. Le, J.S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer, *Sticky central limit theorems on open books*, Ann. Appl. Probab. **23** (2013), 2238–2258.

[3] T.M.W. Nye, X. Tang, G. Weyenberg, and R. Yoshida, *Principal component analysis and the locus of the Fréchet mean in tree space*, Biometrika **104** (2017), 901–922.

[4] M. Owen, and J.S. Provan, *A fast algorithm for computing geodesic distances in tree space*, IEEE/ACM Trans. Comput. Biol. Bioinf. **8** (2011), 2–13.

[5] T. Hotz and S. Huckemann, *Intrinsic means on the circle: Uniqueness, locus and asymptotics*, Annals of the Institute of Statistical Mathematics **67** (1) (2015), 177–193.

[6] B. Eltzner and S. Huckemann, *A Smeary Central Limit Theorem for Manifolds with Application to High Dimensional Spheres*, arXiv:1801.06581

[7] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20(1):1–100, January 2010.

[8] Sungkyu Jung, Ian L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, September 2012.

[9] X. Pennec, *Barycentric Subspace Analysis on Manifolds*, To appear in Annals of Statistics, Institute of Mathematical Statistics. `https://arxiv.org/abs/1607.02833v2`, Oct 2017.

[10] Stefan Sommer. An Infinitesimal Probabilistic Model for Principal Component Analysis of Manifold Valued Data. *arXiv:1801.10341 [cs, math, stat]*, January 2018. arXiv: 1801.10341.

[11] S. Sommer, A. Arnaudon, L. Kuhnel, S. Joshi. Bridge Simulation and Metric Estimation on Landmark Manifolds. arXiv:1705.10943 [cs.CV]

[12] L. Devilliers, S. Allasonnire, A. Trouv and X. Pennec., *Template estimation in computational anatomy: Frchet means in top and quotient spaces are not consistent.* SIAM Journal of Imaging Science, 10(3):1139-1169. (2017).

[13] N. Miolane, S. Holmes, X. Pennec., *Template shape computation: correcting an asymptotic bias.* SIAM Journal of Imaging Science, 10(2):808-844, (2017).

[14] K. Turner S. Mukherjee D. M. Boyer . Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3, 4:310–344, (2014)

[15] L. Crawford, A. Monod, A. X. Chen, S. Mukherjee, R. Rabadán. Functional Data Analysis using a Topological Summary Statistic: the Smooth Euler Characteristic Transform. arXiv:1611.06818 [stat.AP] (2017).

[16] A. Monod, S. Kališnik, J.Á. Patiño-Galindo, L. Crawford. Tropical Sufficient Statistics for Persistent Homology. arXiv:1709.02647 [math.ST] (2017).

Workshop: Statistics for Data with Geometric Structure

Table of Contents