
AMS Short Course in Baltimore, Maryland

AMS Short Course on Geometry and Topology in Statistical Inference

This two-day course will take place on Monday and Tuesday before the meeting actually begins. It is organized by **Sayan Mukherjee**, Duke University.

In the current era of information, large amounts of complex data are routinely generated across science and engineering. There are two fundamental challenges to using this data to understand and model the underlying phenomena: the size of the data and the complexity of the data. Often the objects we would like to model have geometric or topological structure; examples include curves or surfaces such as bones or teeth, positive definite matrices or subspaces that describe variation in phenotypic traits due to genetic variation, or the geometry of multivariate trajectories generated from cellular processes or an attack on a computer network.

Modeling these types of data has motivated the use of ideas from geometry and topology in data analysis and has become more prevalent in statistics, computer science, and mathematics. For example, there is a year-long program at the IMA on Scientific and Engineering Applications of Algebraic Topology and a year-long program at SAMSI on low-dimensional representations of high-dimensional data. Two burgeoning research topics related to geometric and topological data analysis are manifold learning, the idea that high-dimensional data is concentrated on low-dimensional manifolds; and topological data analysis, using topological summaries computed from data to model and understand the underlying structure in the data.

In this short course we will explore how geometry and topology are being used in statistical inference to build models that extract structure from data. The main mathematical and statistical ideas we will develop are stochastic models and analysis using geometric and topological objects. The applications we will look at include geometric and topological analysis in (cancer) systems biology, modeling rankings in social networks, and analysis of games using Hodge theory, subspace, and covariance models in quantitative genetics and statistical genetics.

Each talk will be accessible to a general audience and will contain several open questions and/or suggestions for new directions of research. The talks on the first day will provide an overview of the statistical and computational challenges, respectively, in using geometric and topological analysis. An effort will be made in both talks to highlight how geometry and topology play a key role in stochastic modeling and computing. The talks on the second day will consider some specific topics: the interface of topology and geometry with probability, geometry and topology in cancer systems biology, and applications of Hodge theory in data analysis.

Day One Lectures

(1) **Sayan Mukherjee**, Duke University, *Geometry in statistical inference*: Geometric approaches to data analysis, including manifold learning, subspace inference, factor models, and inferring covariance/positive definite matrices. Applications will be used to highlight methodologies.

(2) **Sayan Mukherjee**, Duke University, *Topology in statistical inference*: Probabilistic perspectives on topological summaries of data such as persistence homology and inference of topological summaries based on the Hodge operator and the Laplacian on forms. Again applications will be used to highlight methodologies.

(3) **Yusu Wang**, Ohio State University, *Computing geometric and topological summaries*: Algorithms for computing geometric and topological summaries of data, including persistence homologies, computational aspects of manifold learning, and distance-based computations in high dimensions.

Day Two Lectures

(4) **Monica Nicolau**, Stanford University, *Geometry and topology in cancer systems biology*: Examples of geometric and topological data analysis in cancer systems biology. Topological data analysis is used to analyze breast cancer transcriptional data and identify a unique subgroup of Estrogen Receptor-positive (ER+) breast cancers with excellent survival prognosis. A geometric approach to high-dimensional data analysis, called disease-specific genomic analysis (DSGA), will be discussed.

(5) **Matthew Kahle**, Ohio State University, *Random geometry and topology*: The geometry and topology induced by random processes, the topology of random clique complexes, random geometric complexes, limit theorems of Betti numbers of random simplicial complexes.

(6) **Lek-Heng Lim**, University of Chicago, *Hodge operator in data analysis*: Applications of discrete Hodge theory on simplicial complexes to problems in game theory, graphics, imaging, learning, ranking, robotics, voting, and sensor networks.

There are separate registration fees to participate in this course. Advance registration (before December 24): Member US\$106, Nonmember US\$155; Student, unemployed, or emeritus US\$54. Onsite registration: Member US\$140; Nonmember US\$185; Student, unemployed, or emeritus US\$75. To register, go to http://jointmathematicsmeetings.org/meetings/national/jmm2014/2160_regfees.