LOW-DIMENSIONAL STRUCTURE IN HIGH-DIMENSIONAL SYSTEMS 2013–2014 SAMSI PROGRAM REPORT, DECEMBER 2014

DIRECTORATE LIAISON: EZRA MILLER

The LDHD program was devoted to the development of methodological, theoretical, and computational treatment of high-dimensional mathematical and statistical models. Possibly limited amounts of available data pose added challenges in high dimensions. The program addressed these challenges by focusing on low-dimensional structures that approximate or encapsulate given high-dimensional data. Cutting edge methods of dimension reduction were brought together from probability and statistics, geometry, topology, and computer science. These techniques included variable selection, graphical modeling, classification, dimension reduction in matrix estimation, empirical processes, and manifold learning. Working groups during the program included theoretical discussions of these tools as well as applications to image and signal analysis, graphs and networks, genetics and genomics, online streaming, and machine learning.

Representative general research topics include

- sparse structures
- regularization techniques
- confidence regions and p-values in high dimensions
- priors favoring concentration of posterior distributions around low-dim solutions
- topological and geometric techniques for data analysis
- biological and computational applications, such as metagenomics.

Specific research foci included: statistical inference for low-dimensional structures; graph and network estimation; variable selection, screening, and multiple testing; graphical models; statistical applications of topology and geometry; dynamical systems; data sketching; low-dimensional representation of genetics and genomics; image and signal analysis; asymptotic geometric analysis; computational aspects; and matrix estimation under complexity constraints.

Contents.

Organizing committee.

1. Program Leaders
   - Florentina Bunea (Cornell, Statistics)
   - Peter Hoff (Washington, Statistics)
   - Chris Holmes (Oxford, Statistics)
   - Peter Kim (Guelph, Math & Stat)
   - Vladimir Koltchinskii (Georgia Tech, Mathematics)
   - John Lafferty (U Chicago, Statistics and Computer Science)
   - Gilad Lerman (Minnesota, Mathematics)
   - Sara van de Geer (ETH Zurich, Statistics)
   - Marten Wegkamp (Cornell, Statistics)
   - Bin Yu (Berkeley, Statistics)
2. Local Scientific Coordinators
   - Sayan Mukherjee (Duke, Stat)
   - Andrew Nobel (UNC, Statistics)
3. Liaison to 2013–2014 IMA program on Applications of Algebraic Topology
   - John Harer (Duke, Math)

LDHD Workshops.

1. Planning Workshop
   - 17 May 2012 at SAMSI
   - A few experts talked for a day about focus topics and potential participants for the program. The Program Leaders gave four half-hour presentations to outline their general vision for the LDHD program. The rest of the day was spent brainstorming to identify subtopics of LDHD suitable for Working Groups during the program; potential participants in the program, including Working Groups and Opening Workshop speakers; events during the LDHD year, including a Summer School and graduate courses
   - Subjects with LDHD represented: machine learning; statistics (including high dimension probability and statistical inference); Bayesian methods; compressed sensing; applied and interdisciplinary topics; mathematical aspects, particularly geometry & topology; biological applications
   - Attendees
     a. Florentina Bunea (Cornell, Stat),
     b. Ingrid Daubechies (Duke, Math),
     c. Barbara Engelhardt (Duke, Genome Sciences),
     d. Subhashis Ghoshal (NC State, Stat),
     e. John Harer (Duke, Math),
     f. Peter Kim (Guelph, Math & Stat),
     g. Vladimir Koltchinskii (Georgia Tech, Math),
     h. John Lafferty (Chicago, Stat),
     i. Gilad Lerman (Minnesota, Math),
     j. Karim Lounici (Georgia Tech, Math),
     k. Steve Marron (UNC, Stat/OR),
     l. Sayan Mukherjee (Duke, Stat),
     m. Hari Narayanan (U Washington, Math),
     n. Andrew Nobel (UNC, Stat/OR),
     o. Alessandro Rinaldo (Carnegie Mellon, Stat),
     p. Justin Romberg (Georgia Tech, Computer Science),
     q. Marten Wegkamp (Cornell, Stat),
     r. Martin Wells (Cornell, Stat),
     s. Rebecca Willett (Duke, Electrical & Computer Engineering),
     t. Yichao Wu (Temple, Stat),
     u. Harrison Zhou (Yale, Stat)
2. Summer School
   - 11–16 August 2013 at SAMSI
   - Organizers
     - Florentina Bunea (Cornell, Stat)

    - Sayan Mukherjee (Duke, Stat)
    - Yichao Wu (was at Temple, Stat when agreed; now at NCSU, Stat)
- Lecturers and topics
    a. Vin de Silva (Pomona, Math)
    – Topological and geometrical structures in data analysis
    b. David Dunson (Duke, Stat)
    – Bayesian learning from big data
    c. Ann Lee (Carnegie Mellon, Stat)
    – Population and familial structure in genetic association studies
    d. Elizabeth Meckes (Case Western, Math)
    – Randomness in geometry and topology: finding order in the chaos
    e. Tong Zhang (Rutgers, Stat)
    – Convex and nonconvex methods for high dimensional sparse estimation
    f. Hua Zhou (NC State, Stat)
    – Genomics and high-dimensional optimization
- Attendees: 81
- Poster session: Monday 12 August

3. Opening Workshop
    - 8–12 September 2013 at SAMSI
    - Organizers
        - Chris Holmes (Oxford, Stat)
        - Vladimir Koltchinskii (Georgia Tech, Math)
        - Gilad Lerman (Minnesota, Math)
        - Sara van de Geer (ETH Zürich, Stat)
    - Tutorial speakers (Sunday 8 September)
        a. Arnak Dalalyan (ENSAE/CREST)
        b. Ming Yuan (Georgia Tech, Math)
        c. Yurii Nesterov (Louvain, Mathematical Engineering)
        d. David Dunson (Duke, Stat)
        e. Anna Gilbert (Michigan, Math)
    - Invited lecturers (Monday to Wednesday): Invited speakers addressed specific research topics relevant to Working Groups in the program
        a. Tony Cai (Penn, Wharton)
        b. Ingrid Daubechies (Duke, Math)
        c. Inderjit Dhillon (UT Austin, Computer Science)
        d. Gilad Lerman (Minnesota, Math)
        e. Liza Levina (Michigan, Stat)
        f. Karim Lounici (Georgia Tech, Math)
        g. Sayan Mukherjee (Duke, Stat)
        h. Boaz Nadler (Weizmann, Computer Science)

  i. Yaniv Plan (Michigan, Math)
  j. Joel Tropp (Cal Tech, Math)
  k. Rebecca Willett (Wisconsin, Electrical & Computer Engineering)
  l. John Wright (Columbia, Electrical Engineering)
  m. Bin Yu (Berkeley, Stat)
  n. Cun-Hui Zhang (Rutgers, Stat)

- Attendees: 167
- Poster session: Monday 9 September
- Working Group formation: Wednesday 11 September
- Initial Working Group meetings: Thursday and Friday, at SAMSI. Each working group identified initial research activities and relevant datasets for the program year. These foci were intended to evolve over the year, and Working Groups have merged or emerged as dictated by research progression.

The Opening Workshop provided an overview of core topics relevant to the LDHD program, which is devoted to the development of methodological, theoretical, and computational treatment of high-dimensional mathematical and statistical models.

4. Mid-program workshop: Topological Data Analysis
   - 3–7 February 2014 at SAMSI
   - Organizers
     - Robert Adler (Technion, Electrical Engineering)
     - Paul Bendich (Duke, Math)
     - John Harer (Duke, Math)
     - Sayan Mukherjee (Duke, Stat)
   - Invited speakers
     a. Yuliy Baryshnikov (U. Illinois, Math)
     b. Paul Bendich (Duke, Math)
     c. Omer Bobrowski (Duke, Math)
     d. Peter Bubenik (Cleveland State, Math)
     e. Frederic Chazal (INRIA, Geometrica)
     f. Harish Chintakunta (NC State, ECE)
     g. Brittany Fasy (Tulane, Computer Science)
     h. Jennifer Gamble (NC State, ECE)
     i. Giseon Heo (U. Alberta, Statistics)
     j. Peter Kim (U. Guelph, Math & Stat)
     k. Fabrizio Lecci (Carnegie Mellon, Statistics)
     l. Lek-Heng Lim (U. Chicago, Statistics)
     m. J.S. Marron (UNC-CH, Statistics)
     n. Facundo Memoli (Ohio State, Math)
     o. Elizabeth Munch (IMA)
     p. Andrew Nobel (UNC-CH, Statistics)

- q. Megan Owen (Lehman College, Computer Science)
- r. Vic Patrangenaru (Florida State, Statistics)
- s. Jose Perea (Duke, Math)
- t. Alessandro Rinaldo (Carnegie Mellon, Statistics)
- u. Katharine Turner (U. Chicago, Math)
- v. Bei Wang (U. Utah, Scientific Computing and Imaging)
- w. Yusu Wang (Ohio State, Computer Science and Engineering)
- x. Larry Wasserman (Carnegie Mellon, Statistics)

- Attendees: 64
- Poster session: Tuesday 4 February

Topological Data Analysis (TDA) began a little more than 20 years ago with the introduction of persistence by Edelsbrunner, Letscher and Zomorodian. Since that time the field has grown significantly. Many strong theorems, numerous algorithms, and pieces of software have been created in that period. More recently, the field has considered stochastic modeling and analysis, which has resulted in interactions with Stastistics, Probability, and Computer Science. This workshop focused on the interaction of Statistics and Probability with this new area of Applied Mathematics. The goals were to familiarize people in both areas with techniques from the other and to understand what future directions are the most promising. Topics will include:

- Statistics and persistent homology
- TDA and inference in dynamical systems and time series
- TDA and shape statistics
- TDA and network models
- Probability models for random topology objects
- Statistical applications of Hodge theory

5. Undergraduate workshop
   - 20–21 February 2014 at SAMSI

   For details of this event, see the E&O (Education and Outreach) portion of the 2013–2014 annual report.

6. Mid-program workshop: Statistical inference in sparse high-dimensional models: theoretical and computational challenges
   - 24–26 February 2014 at SAMSI
   - Organizers
     - Florentina Bunea (Cornell, Stat)
     - Marloes Maathuis (ETH Zürich, Stat)
     - Marten Wegkamp (Cornell, Stat)
   - Invited speakers
     - a. Yannick Baraud (Nice)
     - b. Jacob Bien (Cornell)
     - c. Tony Cai (UPenn)

d. Venkat Chandrasekaran (CalTech)
e. Cristophe Giraud (Paris)
f. Chris Holmes (Oxford)
g. Han Liu (Princeton)
h. George Michailidis (U of Michigan)
i. Andrew Nobel (UNC)
j. Sofia Olhede (UC London)
k. Xiaotong Shen (U Minnesota)
l. Sasha Tsybakov (Paris)
m. Luo Xiao (Biostatistics, Johns Hopkinks, School of Public Health)
n. Cun-Hui Zhang (Rutgers)
o. Helen Zhang (Arizona)
p. Harry Zhou (Yale)

- Attendees: 77
- Poster session: Monday 24 February

This workshop focused on both theoretical and computational developments in high-dimensional statistical models. Of particular interest were models that involve high-dimensional matrix estimation, such as elliptical copula models, graphical and network models, factor models, and functional data. These models are typically parametrized by matrices of reduced complexity, for instance of low rank, low effective rank, with sparse patterns, or some combination of these. The low-complexity assumptions are crucial for the successful implementation and theoretical analysis of such models, especially from limited data.

High-dimensional models with low-dimensional structures are ubiquitous. Rich applications occur in genetics, neuroscience, economics, public health, psychology and sociology. New scientific challenges in these established areas, or in emerging areas such as medical geology or action science, arise on a continual basis, and with them the need to meet them at both computational and theoretical levels.

The workshop brought together researchers in applied, computational, and theoretical statistics, with the goals of (i) identifying pressing scientific open questions that can be answered within the framework of the workshop; (ii) disseminating state of the art results in the area of high dimensional statistical inference; and (iii) identifying open theoretical and computational challenges in this area.

7. Mid-program workshop: Geometric aspects of high-dimensional inference
   - 31 March–2 April 2014 at SAMSI
   - Joint venture between SAMSI and Centre de Recerca Matemàtica (CRM) in Barcelona, Spain; reciprocal CRM–SAMSI event at CRM is later in 2014
   - Organizers
     - Vladimir Koltchinskii (Georgia Tech, Math)
     - Karim Lounici (Georgia Tech, Math)

  - Shahar Mendelson (Technion & Australian National, Math)
  - Alexandre Tsybakov (CREST, Stat)
  - Sara van de Geer (ETH Zurich, Stat)
- Invited speakers
  a. Witold Bednorz (Mathematics, University of Warsaw, Poland)
  b. Peter Bickel (Statistics, Berkeley)
  c. Cristina Butucea (Statistics, University Marne-le-Vallee, France)
  d. Sourav Chatterjee (Mathematics and Statistics, Stanford)
  e. Guillaume Lecue (Mathematics, University Marne-le-Vallee, France)
  f. Gabor Lugosi (Economics, Pompeu Fabra University, Barcelona, Spain)
  g. Zongming Ma (Statistics, UPenn)
  h. Rob Nowak (Electrical and Computer Engineering, Wisconsin)
  i. Grigoris Paouris (Mathematics, Texas A&M)
  j. Alexander Rakhlin (Statistics, UPenn)
  k. Philippe Rigollet (Statistics, Princeton)
  l. Justin Romberg (Electrical Engineering, Georgia Tech)
  m. Roman Vershynin (Mathematics, University of Michigan)
  n. Jon Wellner (Statistics, University of Washington)
  o. Harry Zhou (Statistics, Yale University)
  p. Shuheng Zhou (Statistics, University of Michigan)
- Attendees: 45

The development of methods of statistical inference for high-dimensional data has become a focal point of research in statistics and machine learning in the recent years. One of the crucial problems is to understand how to estimate efficiently a very high-dimensional object under certain "low complexity" constraints that make the estimation possible. In specific settings, "low complexity" could mean, for instance, sparsity of a vector in a high-dimensional space or low-rank properties of a large matrix. The goal is to develop methods that would be adaptive to the underlying "low complexity" structure. Such problems are extremely important both in statistics—for instance, in high-dimensional regression or in covariance matrix estimation—and in a variety of applications, including compressed sensing, collaborative filtering, and quantum state tomography.

Theoretical analysis of methods of high-dimensional inference often relies on deep understanding of the underlying geometry of high-dimensional spaces that leads to highly nontrivial problems of geometric nature. Similar problems have occurred and have been studied in such areas of mathematics as high-dimensional probability, random matrix theory, asymptotic geometric analysis, convex geometry, and additive combinatorics. Some of the tools developed in these areas proved to be extremely useful in high-dimensional statistics; these tools include empirical processes methods, concentration inequalities, and various techniques from the theory of random matrices. There is also great potential for applications of other tools developed in recent years, such as generic chaining bounds for stochastic processes or an emerging theory of log-concave distributions in high-dimensional spaces.

The goal of this workshop was to bring together researchers actively working on the development of high-dimensional inference in statistics, machine learning, compressed sensing, and other related areas with mathematicians who made major contributions to high-dimensional probability and asymptotic geometric analysis in recent years. This provided a great opportunity for fruitful discussions of cutting edge problems in high-dimensional statistics and major advances in our understanding of geometry of high-dimensional spaces.

Topics covered by the workshop:

- low rank matrix estimation;
- sparse recovery and compressed sensing;
- covariance estimation for high-dimensional data;
- model selection and oracle inequalities in high-dimensional statistics;
- statistical inference for log-concave distributions in high dimensions;
- non-asymptotic theory of random matrices;
- concentration inequalities, generic chaining, and empirical processes methods in high-dimensional statistics;
- hypotheses testing for high-dimensional objects.

8. Transition workshop
   - 12–14 May 2014 at NCBC
   - Organizers
     - Peter Kim (U. Guelph, Math & Stat)
     - Hélène Massam (York U., Stat)
     - Marten Wegkamp (Cornell, Stat)
   - Invited speakers
     a. High-dimensional graphical models
        - Elizabeth Gross (North Carolina State, Mathematics)
        - Hélène Massam (York U., Statistics)

     b. Topological methods
        - Bailey Fosdick, SAMSI
        - Ezra Miller (SAMSI & Duke, Mathematics)

     c. Genetics and genomics
        - Wei Sun (University of North Carolina, Biostatistics)
        - Chunhua Xing (AstraZeneca – MedImmune, Biostatistics)

     d. Image analysis, signal analysis, computer vision
        - Oleg Makhnin (New Mexico Inst. of Mining and Technology, Stat)
        - Xu Wang (Minnesota, Mathematics)

     e. Statistical inference for large matrices under complexity constraints
        - Emanuel Ben-David (Columbia, Statistics)

- Yin Xia (University of North Carolina, Statistics)

f. Semi-parametric models
  - Zuofeng Shang (Purdue, Statistics)
  - Xianyang Zhang (Missouri, Statistics)

g. Inference for dimension reduction
  - Yifan Xu (Case Western, Epidemiology & Biostatistics)
  - Lingsong Zhang (Purdue, Statistics)

h. Nonlinear low-dimensional structure in high dimensions for biological data
  - J.S. Marron (UNC-CH, Statistics)
  - Jungsik Noh (Seoul National University, Statistics)

i. Data analysis on Hilbert manifolds and their applications
  - Emil Cornea (University of North Carolina, Biostatistics)
  - Vic Patrangenaru (Florida State, Statistics)

j. Online streaming and sketching
  - Mikhail Belkin (Ohio State University, Mathematics)
  - David Lawlor (SAMSI & Duke, Mathematics)

- Attendees: 46 (on any given day; total distinct participants closer to 75)
- Poster session: Monday 12 May

The goals of the workshop are to

- reunite participants from all active working groups in the LDHD Program;
- report and review the progress of the working groups; and
- foster continuation of the working group research beyond the LDHD Program.

9. CANSSI–SAMSI: Geometric topological and graphical model methods in statistics
   - 22–23 May 2014 at Fields Institute, Toronto, Canada
   - Jointly funded by CANSSI (Canadian Statistical Sciences Institute) and SAMSI, with local support from the Fields Institute
   - Organizers
     - Peter Kim (U. Guelph, Math & Stat)
     - Hélène Massam (York U., Stat)
     - Ezra Miller (SAMSI & Duke, Math)
   - Invited speakers

     a. Syed Ejaz Ahmed (Brock, Mathematics and Statistics)
     b. Emanuel Ben-David (Columbia, Statistics)
     c. Joseph Beyene (McMaster, Biostatistics)
     d. Peter Bubenik (Cleveland State, Mathematics)
     e. David Dunson (Duke, Statistics)
     f. Subhashis Ghosal (North Carolina State, Statistics)
     g. Elizabeth Gross (North Carolina State, Mathematics)
     h. Giseon Heo (University of Alberta, Dentistry and Statistics)
     i. Stephan Huckemann (Göttingen, Stochastiks)

j. Georges Michailidis (Michigan, Statistics)
k. Washington Mio (Florida State, Mathematics)
l. Sayan Mukherjee (Duke, Statistics)
m. Victor Patrangenaru (Florida State, Mathematics)
n. Thanh Mai Pham Ngoc (Paris-Orsay, Mathematics)
o. Bala Rajaratnam (Stanford, Statistics)
p. Elena Villa (Università degli Studi di Milano, Mathematics)

- Attendees:
- Poster session: Thursday 22 May

Massive, high-dimensional data sets, for which traditional methods are inadequate, pose challenges in processing, interpretation and analyses. These challenges have led to increased innovations in statistical methods to deal with the scale and complexity of the data. A fusion of various approaches is required. The purpose of this workshop is to highlight some of the innovations obtained with geometric, topological and graphical model methods and show some of their applications to areas such as bioinformatics, genetics, and neurosciences, to name a few. The kernel of this workshop stems from some of the working groups coming from the 2013–2014 SAMSI LDHD program. Invited speakers will present their methodological advancements with a heavy emphasis on applications.

LDHD Grad course. Geometric and topological summaries of data and inference

1. Course leader, Fall: Sayan Mukherjee (Duke, Stat)
   Synopsis: The course focused on geometric and topological summaries computed from data that are routinely generated across science and engineering. The focus was on modeling objects that have geometric or topological structure. Examples included curves, or surfaces such as bones or teeth, or objects of higher dimension such as 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. Specific topics included the following.
   a. Geometry in statistical inference – Material covered included recent work in machine learning and statistics on the topics of manifold learning, subspace inference, factor models, and inferring covariance/positive definite matrices. Applications were used to highlight methodologies. The focus was on methods used to reduce high-dimensional data to low-dimensional summaries using geometric ideas.
   b. Topology in statistical inference – Material covered focused on probabilistic perspectives on topological summaries such as persistent homology and on inference of topological summaries based on the Hodge operator and the Laplacian on forms. Again, applications were used to highlight methodologies.
   c. Random geometry and topology – Material covered the geometry and topology induced by random processes. Topics included the topology of random clique complexes, random geometric complexes, and limit theorems of Betti numbers of random simplicial complexes.
   d. Applications of the Laplacian operator in data analysis – Material covered the various uses of the Laplacian in data analysis, including manifold learning, spectral clustering, and Cheeger inequalities. More advanced topics included the Hodge operator (or combinatorial Laplacian) and applications to data analysis, including decomposing ranked data into consistent and inconsistent components, inference of structure in social networks, and decomposing games into parts that have Nash equilibria and parts that cycle.

   Prerequisites: Background in calculus and linear algebra and some reasonable foundation in statistics and probability.
   Course Format: The main instructor was Sayan Mukherjee but there were several guest lecturers, with material and instructors paralleling certain of the major themes in the 2013-2014 year-long SAMSI program on Low-Dimensional Structure in High-Dimensional Systems (LDHD).
   Registration for this course processed through the universities:
   - Duke: STA 790.01
   - NCSU: MA 810.001
   - UNC: STOR 892.1

   Attendees:
   1. Jonathan Christensen (Duke, Stat)

2. Christopher Glynn (Duke, Stat)
3. James Johndrow (Duke, Stat)
4. Lizhen Lin (Duke, Stat)
5. Irving Salvatierra (Duke, Econ)
6. Colbert Sesanker (Duke)
7. Jacopo Soriano (Duke, Stat)
8. Hamza Ghadyali (Duke/SAMSI, Math)
9. Wenjing Liao (Duke/SAMSI, Math)
10. Akihiko Nishimura (Duke/SAMSI, Math)
11. Brian St. Thomas (SAMSI & Duke Stat)
12. Michael Benfield (NCSU, Math)
13. Joe Burdis (NCSU, Math)
14. Alireza Dirafzoon (NCSU, Electrical Engineering)
15. Saba Emrani (NCSU, Electrical Engineering)
16. Jennifer Gamble (NCSU, Electrical Engineering)
17. Ian Haywood (NCSU, Math)
18. Shikai Luo (NCSU, Stat)
19. Eun Jeong Min (NCSU, Stat)
20. James Nance (NCSU, Applied Math)
21. Thomas Wentworth (NCSU, Applied Math)
22. Snehalata Huzurbazar (SAMSI, Stat)
23. Stephen Rush (SAMSI & Guelph, Math)
24. Sanvesh Srivastava (SAMSI, Stat)
25. Dane Taylor (SAMSI, Math)
26. Kelly Bodwin (UNC-CH, Stat/OR)
27. Wayne Lee (UNC, Math)
28. John Palowitch (UNC, Stat)
29. Dongqing Yu (UNC-CH, Stat/OR)
30. Qiang Sun (UNC/SAMSI, Biostat)
31. Ye (Eric) Wang (Duke, Stat)
32. Rachel (Rujie) Yin (Duke, Math)
33. Jianling Zhong (Duke, Computational Bio & Bioinfo)

2. Course leader, Spring: Peter Kim (Guelph, Math & Stat)
   Synopsis: This course was a continuation of the fall offering, but the material covered did not necessarily require the previous course as a prerequisite. Part of the aim was to provide some of the necessary background needed for material covered in the three spring LDHD mid-program workshops. Topics:
   a. Persistent Homology – In computational algebraic topology, one attempts to recover qualitative global features of the underlying data – such as connectedness, or the number of holes, or the existence of obstructions to certain constructions

– based upon a 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 can potentially lead to robust statistical procedures. A combinatorial construction converts the data into an object forwhich it is possible to compute the topology. A multi-scale 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.

b. Morse Theory – The geometry of Morse functions can completely characterize the topology of an object by the way in which topological characteristics of sublevel sets change at critical points. Indeed, classical Morse theory tells us that the homotopy type is characterized by attaching a cell, whose dimension is determined by the number of negative eigenvalues of the Hessian at a critical point, to the boundary of the set at the critical point. This indeed is a pathway that connects geometry with topology, and one which also serves as a bridge to statistics.

c. SiZer – In its original form, SIgnificant ZERo crossings of derivatives (SiZer) is a graphical tool that helps one visually understand graphical features of a surface as the scale, resolution, or bandwidth changes. What was mainly a graphical aid turns out to have some deep geometrical structure that is only beginning to be understood and has precise ties with topics (1) and (2). The course examined this connection.

d. Metagenomic Data Analysis – The data come from massively parallel sequencing (MPS), otherwise referred to as high-throughput sequencing, next-generation sequencing, or pyro-sequencing. One very important observation that is repeatedly made when studying the microbiome is that it has the structure of a singular mathematical object embedded in high-dimensional space that represents mathematical novelty, but at the same time it poses major technical challenges. The goal here is to produce meaningful quantitative descriptors particularly as it relates to certain gastrointestinal diseases. Although the statistical methods currently used in microbial ecology are standard for statisticians, the microbiology terminology requires some background. The course provided this background with the purpose of being able to apply topics (1)–(3) to this promising field.

Prerequisites: Background in calculus and linear algebra and some reasonable foundation in statistics and probability.

Course Format: The main instructor will be Peter Kim but there will be several guest lecturers, with material and instructors paralleling some of the workshops in the 2013–2014 year-long LDHD program.

Registration for this course processed through the universities:

- Duke: STA 790.01
- NCSU: TBD
- UNC: STOR-892.1 cross listed with MATH 892.1

Attendees: same as first semester plus

33. Michael Casey (Duke, Math)
34. Siyun Yu (UNC-CH, Stat/OR)
35. Florian Wagner (Duke, Biology/Bioinfo)

Working Groups. Leaders of each working group are listed beneath its title; more information on each group can be found at `http://samsi.info/LDHD`

1. High-dimensional graphical models
   - Hélène Massam
2. Probabilistic modeling on moduli spaces
   - David Dunson (Duke, Stat)
   - Sayan Mukherjee (Duke, Stat)

   Statistical methods for Topological Data Analysis
   - John Harer (Duke, Math)
   - Paul Bendich (Duke, Math)
3. Genetics and genomics
   - Barbara Engelhardt (Duke, Genome Sciences & Policy)
   - Li Ma (Duke, Stat)
4. Image analysis, signal analysis, computer vision
   - Gilad Lerman (Minnesota, Math)
5. Statistical inference for large matrices under complexity constraints
   - Vladimir Koltchinskii (Georgia Tech, Math)
   - Karim Lounici (Georgia Tech, Math)
6. Semi-parametric models
   - Guang Cheng (Purdue, Stat)
7. Inference for dimension reduction
   - Naomi Altman (Penn State, Stat)
   - Andreas Artemiou (Michigan Tech, Math)
8. Nonlinear low-dimensional structure in high dimensions for biological data
   - Peter Kim (Guelph, Math & Stat)
   - Steve Marron (UNC, Stat/OR)
9. Data analysis on Hilbert manifolds and their applications
   - Vic Patrangenaru (Florida State, Stat)
   - Hongtu Zhu (UNC, Biostat)
10. Online streaming and sketching
    - Ilse Ipsen (NC State, Math & SAMSI)
    - David Lawlor (SAMSI)
11. Graduate Students
    - Hamza Ghadyali (Grad Fellow from Duke Math)
    - Max Sommerfeld (grad visitor from Göttingen)

Long-term visitors.

1. Beran, Rudy (UC Davis)
   – 1 month in Spring 2014 (April 2014)
2. Bunea, Florentina (Cornell)
   – 2 weeks in Spring 2014 (Feb 17-28, 2014)
3. Butucea, Cristina (Paris Marne-la-Vallee)
   – 1 week in Spring 2014 (Feb 16-22, 2014)
4. Ceyhan, Elvan (Koç, İstanbul)
   – 10 months (the entire LDHD program)
5. Charyyev, Polat (Koç, İstanbul) — grad student
   – 4 months in Spring 2014
6. Cho, Eungchun (Kentucky State)
   – 1 month in Spring 2014 (Feb 2–Feb 26, 2014)
7. Du, Pang (Virginia Tech) — New Researcher
   – 10 months (the entire LDHD program)
8. Eftekhari, Armin (Colorado School of Mines) — grad student
   – 9 months (the entire LDHD program)
9. Giraud, Christophe (Paris-Sud)
   – 2 weeks in Spring 2014 (February–March)
10. Goes, John (Minnesota) — grad student
    – 9 months (the entire LDHD program)
11. Gupta, Pramod (Nehru Hospital, Chandigarh, India, Biostatistics)
    – 6 months (Spring 2014)
12. Heo, Giseon (Alberta)
    – 9 months (the entire LDHD program)
13. Houdré, Christian (Georgia Tech)
    – 2 weeks in Spring 2014
14. Huckemann, Stephan (Göttingen)
    – 3 weeks in Fall 2013
15. Huroyan, Vahan (Minnesota) — grad student
    – 4 months in Fall 2013
16. Kim, Peter (Guelph)
    – 9 months (the entire LDHD program)
17. Klopp, Olga (Paris, Ouest)
    – 2 weeks in February 2014
18. Koltchiniskii, Vladimir (Georgia Tech)
    – 3 months in Fall 2013
19. Kong, Linglong (Alberta) — New Researcher
    – 4 months in Spring 2014
20. Koo, Ja-Yong (Korea)
    – 6 weeks in Spring 2014

21. Letac, Gérard (Toulouse)
    – 3 months in Fall 2013
22. Lounici, Karim (Georgia Tech)
    – 3 months in Spring 2014
23. Makhnin, Oleg (New Mexico Institute of Mining and Technology)
    – 4 months in Spring 2014
24. Massam, Hélène (York University)
    – 3 months in Fall 2013
25. Mendelson, Shahar (Technion & Australian National, Math)
    – 2 weeks in Fall 2013
26. Noh, Jung-Sik (Korea) — New Researcher
    – 7 months (November–May)
27. Park, Junyong (U Maryland, Baltimore County) — New Researcher
    – 4 months in Fall 2013
28. Patrangenaru, Vic (Florida State)
    – sporadic visits in Fall 2013
29. Pensky, Marianna (Central Florida)
    – 2 weeks in Spring 2014
30. Qi, Alan (Purdue)
    – 2 months in Spring 2014
31. Rush, Stephen (Guelph) — grad student
    – 9 months (the entire LDHD program)
32. Sommerfeld, Max (Göttingen) — grad student
    – 9 months (the entire LDHD program)
33. Wang, Nanwei (York University) — grad student
    – 3 months in Fall 2013
34. Wang, Yishi (UNC Wilmington)
    – 3 months in Spring 2014
35. Wegkamp, Marten (Cornell)
    – 2 weeks in Spring 2014 (Feb 17-28, 2014)
36. Xie, Jichun (Temple) — New Researcher
    – 10 months (the entire LDHD program)
37. Xing, Chuanhua Julia (Boston University)
    – 1 month in September 2013
38. Yuan, Ming (Georgia Tech)
    – 2 weeks in Fall 2013 (Nov. 2013)
39. Zhao, Yichuan (Georgia State)
    – 1 month in Spring 2014

Local affiliated personnel.

1. SAMSI Postdocs

  a. Wenjing Liao (from UC Davis, Math)
     mentors: Ingrid Daubechies and Mauro Maggioni (Duke, Math)
  b. Minh Pham (from Rutgers, Operations Research)
     mentor: Guillermo Sapiro (Duke, Electrical & Computer Engineering)
  c. Sanvesh Srivastava (from Purdue, Stat)
     mentors: Barbara Engelhardt (Duke, Genome Sciences & Policy) and
     David Dunson (Duke, Stat)

2. Faculty Fellows
  a. Subhashis Ghoshal (NCSU Stat)
  b. Tim Kelley (NCSU Math)
  c. Li Ma (Duke Stat)
  d. John Harer (Duke Math)
  e. Yufeng Liu (UNC Stat)
  f. Andrew Nobel (UNC Stat)
  g. Hongtu Zhu (UNC Biostat)

3. Graduate Fellows (with thesis advisor listed)
  a. Hamza Ghadyali, Duke Math (Harer)
  b. Akihiko Nishimura, Duke Math (Mattingly)
  c. Rachel Yin, Duke Math (Daubechies)
  d. Jacopo Soriano, Duke Stat (Ma)
  e. Brian St. Thomas, Duke Stat (Mukherjee)
  f. Shikai Luo, NCSU Stat (Ghoshal)
  g. James Nance, NCSU Math (Kelley)
  h. Thomas Wentworth, NCSU Math (Ipsen)
  i. Kelly Bodwin, UNC Stat/OR (Nobel)
  j. Qiang Sun, UNC Biostat (Zhu, Hongtu)
  k. Dongqing Yu, UNC Stat/OR (Shen, Haipeng)

Final reports of Working Groups. Each report begins on page listed

1. High-dimensional graphical models

Starting in September 2013, we had one working group called "High-dimensional graphical models" (henceforth abbreviated WG1). Sometime in October, a subgroup got interested in a particular topic and formed a second group called "Statistical dimension for graphical model selection" (henceforth abbreviated WG2). These two working groups are within the LDHD program.

WG members. Participants who initially came came to one group only are actually now coming to both groups. The composition of both groups is therefore essentially the same and as follows.

Name	Category	Affiliation
H. Massam	WG leader	York U.
N. Wang	Webmaster, Ph.D. Student	York U.
A. Lenarcic	Post-docs	UNC
L. Gross	Post-docs	NCSU
G. Letac	Emeritus	Paul Sabatier, France
E. Ben-David	Faculty	Columbia U.
Junyong Park	Faculty	U. Maryland, Baltimore County
Ray Falk	Consultant	OptInference LLC

Sandipan Roy, Ph.D. student at Michigan State, regularly joins WG2 by phone.
Yichuan Zhao, faculty member at Georgia State, occasionally joins WG1 by phone.

Topics investigated by WG. At the beginning of the program, the group was very interested in the notion of statistical dimension of a cone as presented by Joel Tropp at the opening workshop. We therefore decided to read the following papers:

- "Living on the edge: A geometric theory of phase transitions in convex optimization" by D. Amelunxen, M. Lotz, M. Mccoy and J. Tropp, (2013), arxiv 1303.6672
- "The convex geometry of linear inverse problems" by V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky, *Found. Comput. Math.* **12** (2012), 805–849.

Notes are posted on the group website for both these papers.

After reading these two papers, the group split into two: WG1 decided to read papers on priors on graphs in the context of high-dimensional graphical model selection. The new group, WG2, decided to work on the following research project: find the statistical dimension of the cone of positive definite matrices with fixed zeros.

Presentation topics in WG meetings focused on the following papers.

"Bayesian structure learning in graphical models" by S. Banerjee and S. Ghosal, (2013), Arxiv: 1309.1754.

"Bayesian graphical Lasso models and efficient posterior computation" by H. Wang, (2012), *Bayesian Analysis*, **7**, 867-886.

"On the prior and posterior distributions used in graphical modelling" by M. Scutari, (2012), *Bayesian Analysis*, Arxiv:1201.4058v2

"Kronecker graphs: an approach to modeling networks" by Leskovec et al. (2010), *J. Machine Learning Research*, 985-1042.

"Sparse Gaussian graphical models with unknown block structure" by B. Marlin and K. Murphy, (2010), *Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.*

"Geometry of maximum likelihood estimation in Gaussian graphical models" by Caroline Uhler, (2012), *The Annals of Statistics*, **40**, 238-261, presentation by Elizabeth Gross.

"Living on the edge: A geometric theory of phase transitions in convex optimization" by D. Amelunxen, M. Lotz, M. Mccoy and J. Tropp, (2013), Arxiv 1303:6672 : another reading of the paper, presenttaion by Emanuel Ben-David.

"The Bayesian Lasso", (2008), *Journal of the American Statistical Association*, **103**, 681-686.

"Graphical denoising", (2014), presentation by Alan Lenarcic.

"Clique-Tree based Graphical Denoising", presentation by Alan Lenarcic.

"Gaussian Ranks Of Undirected Graphs", presentation by Emanuel Ben-David.

" Positivity, graphical model and multivariate dependencies I (theory)", presentation by Bala Rajaratnam.

"Positivity, graphical model and multivariate dependencies II (applications)", presentation by Bala Rajaratnam.

Specific products in progress.

1. *Statistical dimension of the cone of positive definite matrices with fixed zeros* (Elizabeth Gross, Emmanuel Ben-David, Gérard Letac and Hélène Massam)

   Our aim is to compute the statistical dimension of the cone of positive definite matrices with fixed zeros. The parameter space for graphical models Markov with respect to an undirected graph is of this type and finding the statistical dimension of such cones would mean that we can say something about the phase transition of the graphical Lasso. The topic is difficult, much more delicate than we thought it would be. We made some progress: we have a manuscript with several original results that we will present at the "Transition Workshop" at SAMSI, May 12th, 2014. We need to progress further before our work is ready for publication.

Organization of follow-up and related workshops.

1. Peter Kim, Hélène Massam, Ezra Miller, *Geometric Topological and Graphical Model Methods in Statistics*, Fields Institute, Toronto, Canada, 22–23 May 2014. http://www.fields.utoronto.ca/programs/scientific/13-14/modelmethods

2. Statistical Methods for Topological Data Analysis

Founded almost fifteen years ago, topological data analysis (TDA) is rapidly becoming an important field. The original mission was to transform traditional algebraic objects into computational tools that could be used on modern datasets, and many compelling applications were then found. The goal of this WG was to better understand how to apply rigorous statistical methods to one of the main TDA outputs: persistence diagrams. Our intentions were both to create theorems and to discuss and find interesting datasets that would inform the statistical theory.

WG members. Active faculty:

1. Sayan Mukherjee, Duke University (Statistics; Computer Science; Mathematics)
2. Andrew Nobel, UNC (Statistics and Operations Research)
3. *Helene Massam, York Univesity (Statistics)
4. Ezra Miller, Duke University (Mathematics)
5. Peter Bubenik, Cleveland State University (Mathematics)
6. David Dunson, Duke (Statistics)
7. John Harer, Duke (Mathematics; Computer Science; ECE)
8. Stephan Huckemann, Goettingen (Mathematics)
9. *Blair Sullivan, NC State (Computer Science)
10. Washington Mio, Florida State (Mathematics)

Postdocs:

1. *Bailey Fosdick, SAMSI and Duke (Statistics)
2. Omer Bobrowski, Duke (Mathematics)
3. Dane Taylor, SAMSI (CMSS)
4. Harish Chintakunta, NC State (ECE)
5. Paul Bendich, Duke (Mathematics)

Graduate students:

1. *Jennifer Gamble, NC State (ECE)
2. Hamza Ghadyali, Duke (Mathematics)
3. Brain St. Thomas, Duke (Statistics)

Topics investigated by WG.

1. *Statistical Inference with Topological Features* (Bendich, Bobrowski, Dunson, Fosdick, Harer, Mukherjee).

   In addition to the WG members listed, this project also involves Nate Strawn, a Math/ECE postdoc at Duke.

   The general idea here is to regard a persistence diagram as a feature vector of bin counts in R^D, via a dividing-up of the plane into D bins, and then to investigate various Bayesian models for the bin-counts in different toy applications. So far the results have mainly been about the 0-dimensional diagrams of families of functions

$f : I \to R$. Even in this simple context, there are some compelling classification results; for example, in digit recognition, or in the classification of driver behavior from speed functions.

The main product thus far has been a paper written by Bendich, Dunson, Fosdick, Harer, and Strawn. However, Bobrowski and Mukherjee are also working on a related project involving diagrams as point processes and dynamical systems. There are also future plans of collaboration between all the members listed in this section.

2. *Brain Artery Trees and Persistent Homology* (Bendich and Miller)

   This project also involves Steve Marron and Sean Skwerer, faculty and graduate student, respectively, at UNC Stats-OR.

   The idea here is to look at a well-known dataset, from Bullit et. al., that consists of around 100 trees, which are meant to represent the arterial strucuture in the brains of around 100 human subjects. There is a general consensus that there should be some correlation between age and some aspect of how the tree sits within the brain; however, previous analyses have not found terribly high correlations. In this work, a simple 0-dimensional persistence diagram associated to a filtration-by-vertical-height of each brain is performed, resulting in around 100 diagrams. Then a few different methods of transforming these diagras into feature vectors are tried. Astonishingly, one finds signficant (around .55) correlation between statistical products of this analysis and age.

   The main product of this project so far is an about-to-be-submitted paper by Bendich, Marron, Miller, and Skwerer.

3. *Brain Artery Trees and Persistence Landscapes* (Bubenik)

   This project involves Steve Marron. Bubenik and Marron are also analyzing the brain data described above, but with the persistence landscapes instead of individual diagrams.

4. Frechet Means of Persistence Diagrams, a Bayesian Approach (Ghadyali, Fosdick, Muhkerjee)

   The goal is to define the mean of a set of persistence diagrams as a continuous distribution over the space of persistence diagrams “D_p”; this expands on the work of Bendich, Munch, Turner, Mukherjee, Mattingly, and Harer, who had constructed the mean as a discrete distribution over the same space. Given a set of persistence diagrams, we first develop a Bayesian approach for defining a probability distribution over all possible couplings between the diagram. We will attempt to show that the mean as we define it in the paper will have several “nice” properties that a mean should have, including generalizing the standard definition of Frechet mean, and stability.

   The main product of this project will probably be a paper entitled “Posterior Distribution of the Frechet Means of a Finite Set of Persistence Diagrams.”

5. *Mouse Lung Trees and Persistent Homology*(Bendich and Miller)

This project also involves the non-WG members Steve Marron, Sean McLean, and Aaron Park; the latter two are a Professor of Pediatric Surgery at UNC and an undergraduate at Duke, respectively. McLean is interested in understanding congenital diagphrameal hernias (CDH) in human infants, and his lab uses infant mice as a proxy to make progress on this. The goal of this project it to use the tree-based persistent homology techniques described above to gain insight on his data, which comes in the form of lung vessel trees. Ideally, one would like to see similar correlations between various persistence diagrams and covariates like blood pressure or age. Even more importantly, one would like to understand the difference between persistence diagrams associated to mice with CDH and those without.

This project is still very much in its preliminary stages.

6. *Hypergraph Models and Random Zig-zag Persistence*Bendich and Mukherjee

   This project builds off the work in the paper "Geometric Representation of Hypergraphs for Prior Specifcation and Posterior Sampling" by Mukherjee, Simon Lunagomez, and Robert Wolpert. The idea in that paper was to use the geometry of R^D to understand the likelihood of various types of hypergraph-based correlation models. However, there was no clear method in that paper for constructing a meaningful metric between two such models.

   In this project, still very preliminary, Bendich and Mukherjee see each pair of models as living within a giant simplex, and they construct a set of random walks in simplex space between them. Each such walk leads to a random homology zigzag, which in turn gives a diagram. By taking weighted means of these diagrams, with the weights determined by probability, they hope to construct meaningful metrics. The obvious inspiration is the diffusion geometry approach of Mauro Maggionni and others.

7. *Asymptotics of sampling from topologically stratified spaces* (Omer Bobrowski, Ezra Miller, Andrew Nobel)

   Statistical problems that deal with high-dimensional or complex data objects, such as images or phylogenetic trees, often involve sampling from spaces or probability distributions that are singular. The singularities can be described by expressing the space or support as a union of smooth manifolds, of varying dimensions, that intersect in controlled but potentially intricate ways. Both for applications and for abstract mathematical purposes, it is necessary to understand fundamental relations between the effects of global as well as local geometric and topological structures on asymptotic behavior of statistical descriptors. The goal of this particular project is to make precise some senses in which increasing sample size gives rise to more accurate topological information, including statements to the effect that in the limit of infinite sample size, topological information deduced from a sample converges to topological information extracted directly from the theoretical probability distribution.

Specific products in progress.

1. *Statistical Analysis with Topological Features* (Bendich, Dunson, Fosdick, Harer, and Nate Strawn)

   This is the paper alluded to above. Its function will be as an easy-to-understand (since 0-dim persistence does not require any algebra!) survey of persistent homology aimed at statisticians, with several toy applications included to show the power of persistence diagrams as features for classification or regression. A key point made in this work is that TDA need not achieve success by itself; rather, PDs can be incorporated as just one of many features in a valuable classification or regression effort.
2. A persistent homology analysis of brain artery trees (Bendich, Miller, Steve Marron, and Sean Skwerer)

   This is the paper alluded to above. It is almost finished. The aim is to submit it to The Annals of Applied Statistics.
3. Posterior distribution of the Fréchet mean of a finite set of persistence diagrams-Ghadyali, Fosdick, Mukherjee

   This is the paper alluded to above. It is almost finished, but the submission target has not yet been decided.
4. *Persistence and machine learning* (Bendich, Harer)

   In colloboration with non-WG members Nate Strawn, Rann Bar-On, and Jurgen Slaczedek, Bendich and Harer have been developing a software suite that does two main things: computes low-dimensional (0-2) persistence diagrams from high-dimensional point clouds using a novel and fast algorithm based on Discrete Morse Theory, and learns how to interpret these diagrams in a variety of classification and regression contexts. The suite is being beta-tested by 6 undergraduate students from the Summer Undergraduate Research Program associated to the NSF-funded Structure in Complex Data RTG at Duke. To facilitate their use of the suite, Bar-On is developing a series of "labs" that will eventually be incorporated into an extensive users manual for the suite. The anticipated public release of the suite is fall 2014. It will in all probability be in the form of an Arxiv paper with a pointer to a download link. It may or may not result in a publication in a refereed journal, as there are several novel algorithms in the suite.
5. Omer Bobrowski, Kevin McGoff, Ezra Miller, and Andrew Nobel, *Asymptotics of sampling from topologically stratified spaces*, 11 pages, incomplete draft dated 10 May 2014.

Other publications related to work done at SAMSI.

Papers submitted/accepted/published in refereed journals:

1. Paul Bendich, Sang Chin, Jesse Clarke, John deSena, John Harer, Elizabet Munch, Andrew Newman, David Porter, David Rouse, Nate Strawn, and Adam Watkins,

Topological and Statistical Behavior Classifiers for Tracking Applications, submitted to IEEE Trans. on Automatic Control.

2. Brian St. Thomas, Lizhen Lin, Lek-Hing Lim, and Sayan Mukherjee, *Learning Subspaces of Different Dimension*, submitted to Journ. of Amer. Stat. Assoc., Theory and Methods.
3. Sayan Mukherjee and Garvesh Raskutti, *Information Geometry of Mirror Descent*, submitted to IEEE Trans. on Information Theory.

Grant proposals.

1. Bendich and Harer, *BIGDATA: Persistence, Machine Learning, and Motion-Capture Data*, NSF, submitted June 2014.
2. Mukherjee and Mio, *Collaborative Research CDS&E-MSS: Topological Methods for Parsing Shapes and Networks and Modeling Variation in Structure and Function*, awarded May 2014.

Conference and workshop presentations.

1. Steve Marron, *Object-Oriented Data Analysis of Tree Data*, SAMSI LDHD Workshop on Topological Data Analysis, SAMSI, Feburary 3–7, 2014.
2. Ezra Miller, *Asymptotics of sampling from topologically stratified spaces*, 11th Stochastiktage: German Probability and Statistic Days (GPSD), Universität Ulm, Germany, 3–7 March 2014. http://gpsd-ulm2014.de/sections.html
3. Ezra Miller, *Applying persistent homology to brain artery imaging*, Stochastics Colloquium, Universität Göttingen, Germany, 10 March 2014. http://www.stochastik.math.uni-goettingen.de/index.php?id=60&language=en
4. Paul Bendich and Nate Strawn, *Topological Features for Machine Learning*, Data Seminar, Duke University, April 10, 2014.
5. Ezra Miller, *Topology for statistical analysis of brain artery images*, Noyce Learning Conference, Duke University, 11 April 2014. (Audience consisted of high-school science and math teachers.)
6. Ezra Miller, *Applying persistent homology to brain artery and vein imaging*, Algebraic Statistics 2014, Illinois Institute of Technology, Chicago, IL, 20 May 2014. http://mypages.iit.edu/~as2014
7. Bailey Fosdick, *Statistical Inference for Topological Features*, LDHD Transition Workshop, SAMSI, May 12–14, 2014.
8. Ezra Miller, *Topological Analysis of Stratified Geometric Data*, LDHD Transition Workshop, SAMSI, May 12–14, 2014.

3. Genetics and genomics

Genetics and genomics are undergoing a major change due to the availability and affordability of modern high-throughput measurement technologies, coupled to biobank and electronic health records. Large scale projects are producing vast quantities of molecular data on human as well as cancer cells and pathogens. Such data includes DNA, RNA, methylation, metabonomic and proteomic meaurements. Statistical methods are needed to assist scientists in interpreting this high-dimensional data, such as methods for reducing dimensionality, to explore pertinent features and dependencies, and associate variation at the molecular level with multivariate clinical phenotypes or population variation. The mission of the WG is to develop statistical methods for addressing the inferential and computational challenges arising in the analysis of high-dimensional genetics and genomics data.

WG members. Active faculty:

1. Barbara Engelhardt, Duke University (Biostatistics & Bioinformatics), WG leader
2. Li Ma, Duke University (Statistics), WG leader
3. Jun Xie, Purdue University (Statistics)
4. Chuanhua Xing, AstraZeneca

Postdocs:

1. Sanvesh Srivastava, Duke University (Stat), Webmaster
2. Alan Lenarcic, UNC (Stat/OR)

Graduate students:

1. Jacopo Soriano, Duke University (Stat)

Occasional members:

1. Anindya Bhadra, Purdue
2. Alan Qi, Purdue (Stat)
3. Wei Sun, UNC

Topics investigated by WG.

1. *Multivariate high-dimensional mixed and multivariate phenotypes*

   In modern genetic association studies, multiple correlated phenotypes are often measured. Traditionally, association testing is carried out on each individual phenotype separately, but more recently it has been shown that analyzing the phenotypes jointly can substantially enhance statistical power. The challenge is to effectively incorporating the correlation structure in the multivariate phenotype, especially in high-dimensional situations where appropriate sparsity constraints must be imposed.

2. *Joint high-dimensional variable and covariance selection*

   Expression quantitative trait locus (eQTL) analysis is a modern apparatus for investigating gene regulation networks. In such studies both the response (gene expression levels) and the covariates (genetic markers such as SNPs) are high-dimensional. The core statistical problem is then to recover the relevant markers to the various responses (variable selection) while taking into account the covariance structure of the response. Due to the high-dimensionality of both the response and the covariates spaces, two types of sparsity—one on the covariance structure of the response and the other on the models formed by collections of the markers. How to achieve this in a computationally effective way is an important open problem.
3. *Models for functional phenotypes*

   In many genetics data sets, the phenotype is functional. In particular, in some new types of sequencing data such as DNAse-seq data, sequence count profile, that is how the sequencing counts vary over a region of the genome, reflect the nature of the related biological mechanism such as transcriptional factor binding. While there are standard methods for fitting regression with functional phenotypes, it is not known what approaches, such as what basis transformations, are most suited for the genetic contexts. Also important is to understand how to address the multiple testing challenge arising from having to fit such models over a large number of genomic locations that are not independent of each other. A related problem is the functional analysis of variance (ANOVA) that aims at decomposing the variability in functional phenotypes into multiple levels and associating them with covariates.
4. *Methods for effective multiple testing control*

   A central challenge in high-dimensional data analysis is the multiple testing correction. Classical strategies for multiplicity adjustment treat the multiple hypotheses as independent, which can be overly conservative in the presence of dependence. In genomic applications, the hypotheses are related according to the underlying biological mechanisms, and therefore biological knowledge can be incorporated to improve the efficiency of multiple testing correction. A powerful strategy for achieving this goal is graphical modeling.

Specific products in progress.

1. *Expandable factor analysis* (Sanvesh Srivastava and Barbara Engelhardt)

   Novel statistical models for high-dimensional reponse and/or covariates motivated by analyzing methylation data, in particular, identifying associations with covariates of interest (disease status, age, etc.). Manuscript in preparation.
2. *Multi-resolution two-sample comparison through the divide-merge Markov tree* (Jacopo Soriano and Li Ma)

   Develop multi-resolution methods for functional phenotypes and for functional analysis of variance, as well as effective multiple testing correction strategies. These methods are developed in the context of high-throughput genomic sequencing data. Manuscript in preparation.

Conference and workshop presentations.

1. Li Ma, *Multi-resolution two-sample comparison through the divide-merge Markov tree*, 6th International Statistics Forum at Renmin University, Beijing, China, May 2014.
2. Li Ma, *Multi-resolution two-sample comparison through the divide-merge Markov tree*, 2014 International Workshop on Controlling Multiplicity in Statistical Analysis, Shanghai, China, June 2014.

4. Image Analysis, Signal Analysis, and Computer Vision

The main goal of the working group was to create a nurturing atmosphere for people interested in its broad area. We chose very recent papers in the area (often in early arxiv form) and assigned group members to present them, while following the presentation with a detailed discussion. More specifically, the group focused on mathematical problems related to performance analysis of algorithms, convex relaxation methods, structure from motion and fast, online, and parallel implementation of algorithms.

We also discussed recent developments and progress of all participants in this very broad area. However, participants were encouraged to pursue projects independently or in very small groups, while using the whole group as a nurturing environment for discussing initial ideas.

WG members. Active faculty:

1. Gilad Lerman, University of Minnesota (Math), WG leader
2. Oleg Makhnin, New Mexico Institute of Mining and Technology (Math)

Postdocs:

1. *Wenjing Liao, SAMSI and Duke University (Math), WG Webmaster
2. Phan Minh, SAMSI (Stat)
3. Dan Kaslovsky, NSF Postdoctoral Fellow at NIST (recently moved to Seagate Technology)
4. Yuyuan (Lance) Ouyang, University of Florida (Math)

Graduate students:

1. *M. Sylvia Agwang, University of Minnesota (Math)
2. Ulas Ayaz, University of Bonn (Math); recently graduated
3. *John Goes, University of Minnesota (Math)
4. *Alex Gutierrez, University of Minnesota (Math)
5. Vahan Huroyan, University of Minnesota (Math)
6. Tyler Maunu, University of Minnesota (Math)
7. Jeffrey Moulton, University of Minnesota (Math)
8. Bryan Poling, University of Minnesota (Math)
9. Xu Wang, University of Minnesota (Math)
10. *Rachel Yin, Duke University (Math)

Occasional members:

1. *Yi Grace Wang, SAMSI and Duke University (Math)
2. Jason Lee, Stanford University (Math)

Topics investigated by WG.

1. *General Topics Reviewed in the Weekly Meetings* (all members)

   Nonlinear dictionary learning, 3D shape matching, tracking with radial distortion, learning-based hashing, generalization of sparse coding and dictionary learning to non-Euclidean settings, distributed dictionary learning, efficient approximation of data by Gaussians with estimation of their number, texture discrimination via group invariance, Grassmannian-based hashing, Super-resolution via superset selection and pruning, recent theoretical developments in justifying algorithms for modeling data by multiple subspaces, recent developments in studying structure-from-motion, basic topics in online learning, information trade-offs in machine learning, the phase retrieval problem, robust PCA, the MUSIC algorithm for line spectral estimation: stability and super-resolution, learning high-dimensional stochastic systems near manifolds, and compressed sensing of videos.

Specific products in progress.

1. *Robust Stochastic Principal Component Analysis* (John Goes and Gilad Lerman; with non-group members: Teng Zhang and Raman Arora)

   We consider the problem of finding lower dimensional subspaces in the presence of outliers and noise in the online setting. In particular, we extend previous batch formulations of robust PCA to the stochastic setting with minimal storage requirements and runtime complexity. We introduce three novel stochastic approximation algorithms for robust PCA that are extensions of standard algorithms for PCA – the stochastic power method, incremental PCA and online PCA using matrix-exponentiated-gradient (MEG) updates. For robust online PCA we also give a sub-linear convergence guarantee. Our numerical results demonstrate the superiority of the robust online method over the other robust stochastic methods and the advantage of robust methods over their non-robust counterparts in the presence of outliers in artificial and real scenarios.
2. *Distributed Robust PCA* (Vahan Huroyan and Gilad Lerman)

   We consider effective distributed algorithm for robust PCA. The project is still in preliminary stage.
3. *Image warping for the low-rank representation of batches of images* (Oleg Makhnin) The works concerns the extension of the well-known RASL method (Peng et al, 2012) for robust batch image alignment, to work in the situations where the image aligning transformations are non-affine. Applications to the low-rank representation of videos are considered.

Publications of Work Discussed During WG Meeting.

Papers submitted/accepted/published in other refereed venues

1. John Goes, Teng Zhang, Raman Arora and Gilad Lerman. *Robust Stochastic Principal Component Analysis*, AISTATS, 2014
http://jmlr.org/proceedings/papers/v33/goes14.html

Other publications related to work done at SAMSI.
Papers submitted/accepted/published in other refereed venues

1. R. Yin, D. Dunson, B. Cornelis, B. Brown, N. Ocon and I. Daubechies, *Digital Cradle Removal in X-ray Images of Art Paintings*, 5 pages, 2014 (submitted to ICIP 2014)

Non-refereed publications

1. Qian Wu, Handwritten digit recognition using image warping and differential evolution adaptive Metropolis, MS thesis at New Mexico Tech (Math), advisor: Oleg Makhnin.

5. Statist. inference for large matrices under complexity constraints

The purpose of this working group was to discuss important problems related to statistical inference for large matrices under various complexity constraints and to try to make progress on some of them. The problems in question included matrix completion, more general problems of matrix estimation based on linear measurements, covariance matrix estimation, estimation of density matrix in quantum state tomography, matrix recovery in compressed sensing, hypotheses testing for large matrices, etc. The complexity constraints include low-rank properties, sparsity of the matrix in a given dictionary, smoothness assumptions for kernels on weighted graphs or manifolds, and various combinations of these complexity assumptions. Analysis of these problems relies on a variety of mathematical tools in high-dimensional probability and asymptotic geometric analysis developed in recent years. In particular, the methods of nonasymptotic theory of random matrices are increasingly important in matrix estimation. Models based on log-concave distributions in high-dimensional spaces could also be of importance (in addition to more classical Gaussian and subgaussian models).

The following more specific topics were subjects of discussions or joint research projects among Working Group members:

- matrix completion under structural assumptions;
- high-dimensional covariance estimation, sparse PCA, PCA for functional data;
- hypotheses testing for large matrices;
- matrix valued time series;
- matrix estimation in statistical problems for networks/graphs;
- noisy matrix decomposition;
- estimation of density matrix in quantum state tomography;
- concentration inequalities for random matrices;
- bounding singular values of large random matrices (e.g., of sample covariance);
- sparsity constraints in matrix and function estimation.

WG members. Active faculty:

1. Florentina Bunea, Professor, Stat, Cornell University.
2. Vladimir Koltchinskii, Professor, Math, Georgia Institute of Technology
3. Karim Lounici, Assistant Professor, Math, Georgia Institute of Technology
4. Marten Wegkamp, Professor of Math and of Stat, Cornell University

Postdocs:

1. Webmaster: Stanislav Minsker, Postdoc, Mathematics, Duke University
2. Junming Yin, Machine Learning and Computational Biology, Carnegie Mellon
3. Wenjing Liao, SAMSI and Duke Math
4. Lingzhou Xue, Operations Research and Financial Engineering, Princeton
5. Xi Chen, Theory of Computing, UC Berkeley
6. Yaniv Plan, Math, U. Michigan.

Graduate students:

1. Dong Xia, School of Mathematics, Georgia Institute of Technology
2. Qiang Sun, Department of Biostatistics, UNC Chapel Hill
3. Jason D. Lee, Computational and Mathematical Engineering, Stanford University
4. Kelly Bodwin, Statistics and Operations Research, UNC Chapel Hill
5. Martina Mincheva, Operations Research and Financial Engineering, Princeton
6. Shujiao Huang, Department of Mathematics, Georgia Southern University
7. Sumanta Basu, Department of Statistics, University of Michigan
8. Jing Ma, Department of Statistics, University of Michigan
9. Wei Sun, Department of Statistics, Purdue University
10. Yuekai Sun, Computational and Mathematical Engineering, Stanford

Other members:

1. Emanuel Ben-David, Assistant Professor, Statistics, Columbia (non-active)
2. Dan Yang, Assistant Professor, Statistics and Biostatistics, Rutgers
3. Jichun Xie, Assistant Professor, Statistics, Temple
4. Junyong Park, Associate Professor, Math & Stat, U. Maryland, Baltimore County
5. Lingsong Zhang, Assistant Professor, Statistics, Purdue
6. Marianna Pensky, Professor, Mathematics, U. Central Florida.
7. Yichuan Zhao, Professor, Mathematics and Statistics, Georgia State
8. Ming Yuan, Professor, Statistics, U. Wisconsin-Madison.
9. Raymond Falk, OptInference LLC.
10. Yin Xia, Assistant Professor, Stat/Operations Research, UNC-Chapel Hill
11. Xingye Qiao, Assistant Professor, Math, Binghamton U.
12. Yimin Xiao, Professor, Statistics and Probability, Michigan State

Topics investigated by WG.

1. New asymptotics of PCA in high dimensions (Koltchinskii, Lounici (Georgia Tech));

   In this project, the goal is to develop new concentration bounds and asymptotic results for eigen-projectors of sample covariance operators in high-dimensional and infinite-dimensional problems. The focus has been primarily on the Gaussian case. The underlying assumption is that the sample size and the effective rank of the true covariance operator are simultaneously large. This framework covers, in particular, the case of so called spiked covariance models intensively studied in the literature. The results obtain so far include new moment bounds and concentration inequalities for the operator norm error of the sample covariance in terms of the effective rank and asymptotic normality results for bilinear forms of eigen-projectors as well as for their Hilbert–Schmidt norm errors.
2. Robust matrix completion (Klopp (Paris-Nanterre), Lounici (Georgia Tech), Tsybakov (CREST));

   In this ongoing project, we study the problem of robust matrix completion that can formulated as follows. We observe noisy entries of $A_0 = L_0 + S_0$ where L_0 is an unknown low-rank matrix and S_0 corresponds to some deterministic corruptions. We wish to recover L_0, but only observe a few entries of L_0 and, among those, a fraction happens to be corrupted by S_0. Of course, we do not know which entries are corrupted. It is particularly relevant in recommender systems applications where malicious users try to manipulate the prediction of matrix completion algorithms by introducing spurious perturbations S_0. It is known that nuclear norm minimization can fail dramatically even if S_0 contains only a single nonzero column. Hence the need for new techniques robust to the presence of corruptions S_0. We propose a robustified version of nuclear norm minimization and prove its statistical optimality under mild conditions on L_0, S_0 and the observation design.

3. Confidence Bands in quantum homodyne tomography with noisy data (Lounici (Georgia Tech), Meziani (Paris-Dauphine));

   Our goal is to derive a minimax adaptive confidence band for the Wigner function W_ρ of a quantum system in quantum homodyne tomography. We propose a construction based on a kernel density deconvolution estimator of W_ρ and established the minimax optimality (up to logs) of this construction on a class of supersmooth Wigner functions.
4. Sparsity in functional regression (Koltchinskii (Georgia Tech), Minsker (Duke));

   This project focuses on the development of L_1-penalization method (LASSO) in the context of functional linear models under an assumption that the "slope" parameter of the model is a "saprse" function in the sense that it can be well approximated by a sum of well separated "spikes". The results obtained so far include oracle inequalities for functional LASSO in the case of subgaussian design showing how the L_2-error of the LASSO estimator depends on the underlying sparsity and other parameters of the problem.
5. Optimal estimation of density matrices in quantum state tomograpy (Koltchinskii, Dong Xia (Georgia Tech));

   The goal of this project is to show that recently developed methods of estimation of large low rank density matrices of quantum systems based on nuclear norm penalization and some other techniques (such as von Neumann entropy penalization) are adaptive in a broad sense. In particular, such methods provide optimal error rates in the whole scale of Schatten p-norm distances as well as in noncommutative Hellinger and Kullback-Leibler distances.
6. Exponential bounds for the smallest singular value of large random matrices in the case of heavy tails (Koltchinskii (Georgia Tech), Mendelson (Technion and ANU));

   The purpose of this project is to obtain sharp exponential bounds on the smallest singular values of large random matrices with i.i.d. rows under weak moment assumptions on the linear forms of the row vectors. Equivalently, the problem is about obtaining sharp exponential bounds on the smallest eigenvalue of the sample covariance matrix in high dimensions. Using several tools from empirical processes theory, a surprisingly simple approach to this problem (usually considered rather hard) has been developed.

7. Adaptive classification using Gaussian copulas (Marten Wegkamp, Yue Zhao (Cornell));

   In this ongoing project, we consider a two-class classification problem. Each class has the same (unknown) Gaussian copula, that is, the same dependence structure, while difference between the classes are expressed by a few different marginal distributions. This study extends the highly studied situation of Gaussian conditional densities with the same covariance matrix, but different means. The latter is oftentimes not realistic and the former is much more general, while retaining the interpretability of the probabilistic underlying model.

8. Analysis of elliptical copula correlation factor model with Kendall's tau (Marten Wegkamp, Yue Zhao (Cornell));

   In this project, we study a factor model for the correlation matrix $\Sigma \in \mathbb{R}^{d\times d}$ of an elliptical copula. The correlations are connected to Kendall's tau and a natural estimation procedure is to plug-in Kendall's tau statistics. The resulting matrix $\widehat{\Sigma}$ can be viewed as a preliminary estimator of Σ and we obtain sharp bound on the operator norm of $\widehat{\Sigma} - \Sigma$. We propose a refined estimator $\widetilde{\Sigma}$ of Σ by fitting a low-rank matrix plus a diagonal matrix to $\widehat{\Sigma}$ using least squares with a nuclear norm penalty on the low-rank matrix. We obtain finite sample oracle inequalities for $\widetilde{\Sigma}$. In addition, we present two estimators based on suitably truncated eigendecompositions of $\widehat{\Sigma}$, one for an elementary factor model and the other for the regime where d is proportional to the sample size.

9. Asymptotic total variation tests for copulas (Marten Wegkamp (Cornell), J.-D. Fermanian (CREST) , D. Radulovic (Florida Atlantic University));

   In this project, we propose a new platform of goodness-of-fit tests for copulas, based on empirical copula processes and nonparametric bootstrap counterparts. The standard Kolmogorov-Smirnov type test for copulas that takes the supremum of the empirical copula process indexed by orthants is extended by test statistics based on the empirical copula process indexed by families of L(n) disjoint boxes, with L(n) slowly tending to infinity. Although the underlying empirical process does not converge, the critical values of our new test statistics can be consistently estimated by nonparametric bootstrap techniques, under simple or composite null assumptions. We implemented a particular example of these tests and our simulations confirm that the power of the new procedure is oftentimes higher than the power of the standard Kolmogorov-Smirnov or the Cramer-von Mises tests for copulas.

10. Analysis of PCA and fPCA based on the sample covariance matrix of reduced effective rank population matrices (Florentina Bunea (Cornell) and Luo Xiao (Johns Hopkins));

In this project, we investigate theoretically the merits and limitations of the scree plot method routinely used in PCA, when data comes from a distribution with covariance matrix that has reduced effective rank. We focus on accurate spectral jump detection and on the estimation of the number of sample eigenvalues and eigenvectors that are optimally close to their theoretical counterpart. The framework of reduced effective rank covariance matrices allows immediate extensions of this analysis to functional PCA, for data generated from certain classes of Gaussian processes.

11. A convex optimization approach to the estimation of banded covariance and inverse covariance matrices (Jacob Bien (Cornell), Florentina Bunea (Cornell) and Luo Xiao (Johns Hopkins));

    We employ a version tailored to matrix estimation of the hierarchical group lasso, with variable weights, that renders naturally banded estimators. We study theoretically the optimality of this estimator with respect to: support recovery and minimax optimal convergence rates in operator norm and Frobenius norm.
12. Several other ongoing projects include the following:
    - development of methods of covariance estimation for functional data (Koltchinskii, Lounici (Georgia Tech), Tsybakov (CREST));
    - estimation of large matrices under simultaneous low rank and sparsity constraints (Lounici, Rangel (Georgia tech));
    - development of optimal testing procedures to detect a change in the covariance structure between two independent populations (Lounici (Georgia Tech));
    - hypotheses testing problems for large matrices (Koltchinskii (Georgia Tech), Ming Yuan (Wisconsin));

Specific products in progress.

1. V. Koltchinskii and K. Lounici, Concentration inequalities and moment bounds for sample covariance operators. arXiv:math.PR/1405.2468

Other publications related to work done at SAMSI.

Papers submitted/accepted/published in refereed journals:

1. V. Koltchinskii and S. Mendelson, Bounding the smallest singular value of a random matrix without concentration, 2013, submitted, arXiv:math.PR/1312.3580
2. V. Koltchinskii and S. Minsker, L_1-Penalization in Functional Linear Regression with Subgaussian Design, submitted, 2013. arXiv:math.ST/1307.8137
3. E. Arias-Castro and K. Lounici, Estimation and variable selection with exponential weights. *Electronic Journal of Statistics*, Volume 8 (2014), 328–354.

Organization of follow-up and related workshops.

1. Sebastien Bubeck (Princeton University), Nicolò Cesa-Bianchi (Università degli Studi di Milano), Gàbor Lugosi (Universitat Pompeu Fabra), Alexander Rakhlin

(University of Pennsylvania), *CRM–SAMSI: Workshop on the Mathematical Foundations of Learning Theory*, 17–19 June 2014 at Centre de Recerca Matemàtica (CRM), Barcelona. http://stat.wharton.upenn.edu/ rakhlin/crm/?q=node/8

Learning theory is a field that lies at the intersection of probability, statistics, computer science, and optimization. This mathematical theory is concerned with theoretical guarantees for machine learning algorithms. Over the last decade the statistical learning approach has been successfully applied to many problems of interest in machine learning, such as bioinformatics, computer vision, speech processing, robotics, and information retrieval. This success story crucially relied on a strong mathematical foundation. Over the past several decades, learning theory has studied inherent complexities of learning problems, both from the statistical and computational points of view. The dialogue between computation and statistics has been key to the enormous advances and growth in the field. The main goal of this workshop is to impulse further these advances by bringing together experts from fields spanning the full broad range of modern learning theory.

6. Semi-parametric models

Parametric modeling provides a convenient inferential setting but is very restrictive. On the other hand, in high-dimensional data, nonparametric models that impose only loose constraints, such as smoothness, require very large sample sizes for convergence of estimators and good inferential performance. Semi-parametric methods allow modeling of some parameters parametrically and others nonparametrically, often improving interpretability and performance of the nonparametric portion without sacrificing efficiency of the parametric portion. This group will explore joint modeling, estimation and inference of the parametric and nonparametric parts for the massive data.

WG members. Active faculty:

1. Guang Cheng, Purdue University (Statistics)
2. David Dunson, Duke University (Statistics)
3. Xingye Qiao, Binghamton University (Mathematics)
4. Zuofeng Shang, Purdue University (Statistics)
5. Xianyang Zhang, Univ. of Missouri – Columbia (Statistics)

Postdocs:

1. Weining Shen (M.D. Anderson Cancer Center)

Graduate students:

1. Wei Sun, Purdue University (Statistics)
2. *Zhuqing Yu, Purdue University (Statistics)
3. *Meimei Liu, Purdue University (Statistics)
4. Yun Yang, Duke University (Statistics)

Topics investigated by WG.

1. *Nonparametric Inference on Functional Data* (Zuofeng Shang, Guang Cheng)

   We propose a regularization approach in making nonparametric inference for generalized functional linear models. In a unified framework, we construct asymptotically valid confidence intervals for regression mean, prediction intervals for future responses, and various statistical procedures for hypothesis testing. In particular, our testing procedure is adaptive to the smoothness of the slope function and covariance function. Despite the generality, our procedure is easy to implement. Numerical examples are provided to demonstrate the empirical advantages over the competing methods.
2. *Bootstrapping High Dimensional Dependent Data* (Xianyang Zhang, Guang Cheng)

   Abstract We focus on the problem of conducting inference for high dimensional weakly dependent time series. Motivated by the applications in modern high dimensional inference, we derive a Gaussian approximation result for the maximum of a sum of weakly dependent vectors using Stein's method, where the dimension of the vectors is allowed to be exponentially larger than the sample size. Our result reveals an interesting phenomenon arising from the interplay between the dependence and dimensionality: the more dependent of the data vector, the slower diverging rate of the dimension is allowed for obtaining valid statistical inference. A type of dimension-free dependence structure is derived as a by-product. Building on the Gaussian approximation result, we propose a blockwise multiplier (Wild) bootstrap that is able to capture the dependence between and within the data vectors and thus provides high-quality distributional approximation to the distribution of the maximum of vector sum in the high dimensional context

3. *Semiparametric Objective Prior* (Yun Yang, Guang Cheng, David Dunson)

   Semiparametric Bernstein-von Mises Theorem has been successfully developed by Bickel and Kleijn (2012) in a general setup among others. This talk mainly focuses on its second order extension with an attempt to figure out the influence of nonparametric Bayesian prior on the semiparametric inference. Such results provide theoretical insights in constructing objective prior in a general semiparametric setup, i.e., so-called semiparametric objective prior.

Specific products in progress. Verbatim the same as “Topics investigated by WG” section, immediately above.

7. Inference for dimension reduction

This working group focused on methodology for dimension reduction, particularly estimating the dimension of the low-dimensional structure and/or number of clusters.

WG members. Active faculty:

1. *Naomi Altman, Penn State University (Stat), WG leader
2. Kossi Edoh, North Carolina Agricultural & Technical State University (Math)
3. Xingye Qiao, SUNY Binghamton (Math)
4. Yifan Xu, Case Western University (Epidemiology and Biostat)
5. Lingsong Zhang, Purdue (Stat)

Graduate students:

1. Wei Luo, Penn State (Statistics)

Occasional members:

1. Andreas Artemiou, Cardiff U. (Stat)
2. John Leichty, Penn State (Marketing)
3. Charles Paulson, Puffinware

Over 50 people signed up for the WG and attended introductory lectures in the first four weeks. A list can be found on the WG webpage.

Topics investigated by WG.

1. *Nonnegative Matrix Factorization* (Altman, Edoh,Luo,Paulson,Qiao,Xu, Zhang)

   We considered basic algorithms for nonnegative matrix factorization including Latent Dirichilet Allocation as a special case, the Brunet method and L2 methods, as well as the use of nonnegative matrix factorization as a clustering method. In this latter formulation, the number of clusters is the same as the dimension of the underlying basis. A number of WG members were already working on other methods for determining the dimension. Our current focus is on methodology for determining dimension, and methods for placing an error envelope on the space spanned by the columns of the basis matrix.

Conference and workshop presentations.

1. Yifan Xu,*Robust Nonnegative Matrix Factorization: Modern Dimension Reduction Procedure for Big Noisy Data Set*, LDHD Transition Workshop, Research Triangle Park, 12–14 May 2014.
2. Lingsong Zhang,*Nested Nonnegative Cone Analysis*, LDHD Transition Workshop, Research Triangle Park, 12–14 May 2014.
3. Kossi Edoh, *Sufficient Dimension Reduction*, LDHD Transition Workshop, Research Triangle Park, 12–14 May 2014.

8. Nonlinear low-dim structures in high dim for biological data

Many problems in biology, particularly at the molecular level, involve very high dimensions. Traditional methods, such as principal component analysis, provide an important first step toward understanding the underlying lower dimensional structure. Nevertheless, finding even lower-dimensional and possibly nonlinear structures requires more qualitative procedures. Some possible candidates, among others, include geodesic principal component analysis, SiZer analysis, and persistent homology. The aim of this Working Group was to describe and understand the nature of certain biological data coming from areas such as brain artery tree networks, gene networks, biomechanical motion data, and metagenomics, to name a few. Its members explored and formulated strategies to best apply these qualitative procedures. In addition, they discussed a recent insight into a connection between SiZer analysis and persistent homology.

WG members. Active faculty:

1. Peter Bubenik, Cleveland State (Mathematics)
2. *Giseon Heo, University of Alberta (Dentistry & Mathematical Sciences)
3. Yuan Jiang, Oregon State University (Statistics)
4. Peter Kim (leader), University of Guelph (Mathematics and Statistics)
5. Steve Marron, University of North Carolina (Statistics and Operations Research)
6. Junyong Park, University of Maryland (Mathematics and Statistics)
7. Victor Patrangenaru, Florida State University (Statistics)
8. Washington Mio, Florida State University (Mathematics)

Postdocs:

1. Jungsik Noh, SAMSI

Graduate students:

1. *Kelly Bodwin, University of North Carolina (Statistics and Operations Research)
2. Diego Diaz, Florida State University (Mathematics)
3. John Fedrouk, University of Alberta (Mathematical Sciences)
4. Pavel Petro, University of Alberta (Mathematical Sciences)
5. *Saadia Pinky, University of Alberta (Mathematical Sciences)
6. Stephen Rush, University of Guelph (Mathematics and Statistics)
7. Max Sommerfield, Goettingen University (Stochastiks)

Occasional members:

1. Rudy Beran, University of California Davis (Statistics)
2. Pang Du, Virginia Tech University (Statistics)
3. Stephan Huckemann, Goettingen University (Stochastiks)
4. Ja Yong Koo, Korea University (Statistics)

Topics investigated by WG.

1. *Computational Topology*

   We examined the role of computational topology as it would apply genomic data. The phylogenetic tree which is very important in computational biology has a very uninteresting topological structure in that it is a one-dimensional contractable manifold. Instead we looked at sequence data directly and used the various metrics associated with comparing DNA nucleotides and found that by constructing a Ripps complex based on the metric, a Betti-1 structure can be observed. This raises a profound issue in that perhaps the phylogenetic tree is not the best representation.
2. *Metagenomics*

   In conjunction with the above an examination into the DNA sequence data was investigated including the collection process. Topics covered include rarefaction curves which include the diversity of different bacterial groups, diproperm which investigates different multivariate patterns, isomap and multi-dimensional scaling, as well as the Dirichlet-Multinomial regression. We applied these methods to irritable bowel syndrome (IBS) data as well as *Clostridium difficile* data.
3. *SiZer*

   Significant zeros (SiZer) was a concept that was developed in the local area at about the same time that Persistent Homology was being developed, also in the area. In a certain sense there great similarities between the two via Morse theory. An investigation into the connection was made and some interesting findings were found. The results will be reported in the Transition Workshop.

Specific products in progress.

1. *The effect of fecal microbiota transplantation on the gut microbiome for recurrent* Clostridium difficile *infection* (WG members to be determined)

   Nineteen patients had their 16S rRNA gene sequenced prior to receiving their initial FMT (pre-FMT), followed by one after they received a final FMT (post-FMT). The data comes from patients who contracted CDI, were recurrent and/or refractory and was treated withFMT by the first author, over the period 2008-2012. There were 94 cases and from this patient pool 19 patients along with 7 donors had their 16S rRNA gene sequenced. Comparing the pre-FMT to the post-FMT sequenced data, a significant increase in diversity occured. Furthermore, in comparison to the donors, there was no significant difference between the donors and post-FMT patients, although the pre-FMT patients were significantly lower. The main Phyla concentrations consisted of Firmicutes, Proteobactaria, Bacteroidetes, and Actinobacteria which accounted for at least 75% for each patient. Of those a significant increase in Bacteroidetes occured, and when compared to the pre-treatment regime, a negative correlation to metronidazle was statistically significant. The novelty of this result quantified verification that Bacteriodetes production is negatively associated with antibiotic usage. The metagenomic information reveals that a significant increase in diversity has occurred after an FMT(s).

2. *Connection between SiZer and persistent homology* (WG members TBD)

   Topological data analysis (TDA) is a flexible methodology for understanding nonlinear low-dimensional structures in high-dimensional data. One of TDA methods is to construct a Čech filtration from data points and to compute persistent homology of the filtration. In this paper, we present a connection between the Čech filtration and a sequence of super-level sets of Gaussian kernel density estimates (KDE) indexed by the bandwidth. The two filtrations are shown to be nested. In particular, we provide a functional relation between the index of Čech filtration and the bandwidth of KDE. By doing so, we propose a new method for calculating topological features of the support of population density, which utilizes the number of critical points of Morse index k based on Morse theory.
3. Software: *Gauss filtration* (WG members to be determined)

   A corresponding `R` package is being pursued.
4. Database: *C. difficile* database (WG members to be determined)

   A release of the patient database.

Other publications related to work done at SAMSI.

Papers submitted/accepted/published in refereed journals:

1. Lee, Belanger, Kassam, Smieja, Higgins, Broukhanski, Kim, *The outcome and long-term follow-up of 94 patients with recurrent and refractory* Clostridium difficile *infection using single to multiple fecal microbiota transplantation via retention enema*, Eur J Clin Microbiol Infect Dis, 4 pp., 2014. doi:10.1007/s10096-014-2088-9
2. Rudy Beran, *Nonparametric estimation of trend in directional data*, submitted. http://arxiv.org/abs/1412.2315

Organization of follow-up and related workshops.

1. Peter Kim, Hélène Massam, Ezra Miller, *Geometric Topological and Graphical Model Methods in Statistics*, Fields Institute, Toronto, Canada, 22–23 May 2014. http://www.fields.utoronto.ca/programs/scientific/13-14/modelmethods

Conference and workshop presentations.

1. Steve Marron, *Significance in scale space*, LDHD Transition Workshop, SAMSI, 13 May 2014.
2. Jungsik Noh, *Linkage between Topological Data Analysis and kernel density estimation*, LDHD Transition Workshop, SAMSI, 13 May 2014.

9. Data analysis on Hilbert manifolds and their applications

At the opening workshop in Fall 2013, out of conversations between potential group participants, including Lizhen Lin and Emil Cornea, a presentation was prepared by Vic Patrangenaru and Hongtu Zhu, to inform other LDHD people on the goals of our group. Later on, this presentation was posted on the WG website and discussed in our first formal meetings at SAMSI.

Our WG has been motivated by the idea of developing statistical tools for data objects that lie on infinite dimensional manifolds and naturally occur in science. Such objects include, for example, functional data seen as functions of a point on a sphere, such as *temperature at a location on Earth's surface*; domain contours such as contours of midsections of a symmetric human body part; random vector fields on the surface of the planet generated by masses of air flow; and so on. Because of the difficulties in handling an infinite number of dimensions, studies reverted often to statistical tools on finite dimensional objects, leading to new dimension reduction inferential methods.

Data Analysis on Manifolds (DAM) is about two decades years old; see [1]. Early precursors of DAM were applied directional data analysis [2] and shape data analysis [3, 4, 5]. DAM started in the nineties with inference for *extrinsic means* [6, 7], function estimation on submanifolds on the Euclidean Space [8]. The first nonparametric results on arbitrary manifolds are due to Patrangenaru [9], also with Bhattacharya [10, 11], who noticed that the CLT can be extended from numerical or functional spaces, to a CLT for the so called *Fréchet means* on sample spaces that admit tangent bundle, therefore have a differentiable structure on them.

An important period in DAM history was the *Analysis of Object Data* program at SAMSI in 2010–2011, with WG's such as "Geometric Correspondence", "Trees" and "Data Analysis on Sample Spaces with a Manifold Stratification". The domain advanced further in 2012, due to an MBI Workshop of *Statistics, Geometry, and Combinatorics on Stratified Spaces Arising from Biological Problems*, and a meeting in Denmark on *Geometry and Statistics.* New methodologies, focusing mainly on the high, or even infinitely dimensional manifolds are in progress, with recent results obtained in this Working Group and the one on "Nonlinear low-dimensional structures in high-dimensions for biological data". The topics that were covered during the year included testing of neighborhood hypotheses in Hilbert manifolds, nonparametric regression with a manifold response, advances in shape analysis, and analysis of nonlinear modes of variation in functional data such as registration problems.

Statistics is ultimately the science of data analysis, so examples from there are crucial. The need for DAM is primarily motivated by the type of data available:

- numerical data
- astronomy and cosmology data: galaxies, stars, and planetary orbit data
- spatial statistics: temperature, snow, other functions measured across the globe
- vector fields of wind velocities on the Earth surface

- geology: paleomagnetic data, plate tectonics, volcanos
- morphometric data
- protein and DNA structures
- medical imaging, including CT and MRI (e.g. angiography: artery structure)
- satellite or aerial imaging
- digital camera imaging data,

to name just a few. Such examples of data were analyzed in our WG meetings using a Hilbert manifolds approach.

References

[1] R. N. Bhattacharya and V. Patrangenaru (2014). Rejoinder of Discussion paper "Statistics on Manifolds and Landmarks Based Image Analysis: A Nonparametric Theory with Applications." *Journal of Statistical Planning and Inference.* **145**, 42–48.

[2] Nicholas I. Fisher, Peter Hall, Bing-Yi Jing, and Andrew T. A. Wood, "Improved Pivotal Methods for Constructing Confidence Regions with Directional Data", Journal of the American Statistical Association **91** (1996), no. 435, 1062–1070. doi:10.1080/01621459.1996.10476976

[3] Kendall, David G., "Shape manifolds, Procrustean metrics, and complex projective spaces", Bull. London Math. Soc. **16** (1984), no. 2, 81–121.

[4] Dryden, I. L. and Mardia, K. V., "Size and shape analysis of landmark data", Biometrika **79** (1992), no. 1, 57–68.

[5] Kent, John T., *New directions in shape analysis*, in "The art of statistical science", 115–127, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., Wiley, Chichester, 1992.

[6] Hendriks, Harrie and Landsman, Zinoviy, *Asymptotic behavior of sample mean location for manifolds*, Statist. Probab. Lett. **26** (1996), no. 2, 169–178.

[7] Hendriks, Harrie and Landsman, Zinoviy, *Mean location and sample mean location on manifolds: asymptotics, tests, confidence regions*, J. Multivariate Anal. **67** (1998), no. 2, 227–243.

[8] Hendriks, Harrie Sur le cut-locus d'une sous-variété de l'espace euclidien. Négligeabilité. (French) [On the cut-locus of a submanifold of Euclidean space. Negligibility] C. R. Acad. Sci. Paris Sér. I Math. **315** (1992), no. 12, 1275–1277.

[9] Patrangenaru, Victor, *Asymptotic statistics on manifolds and their applications*, Ph.D. Thesis, Indiana University, 1998.

[10] Bhattacharya, Rabi and Patrangenaru, Vic, *Large sample theory of intrinsic and extrinsic sample means on manifolds. I*, Ann. Statist. **31** (2003), no. 1, 1–29.

[11] Bhattacharya, Rabi and Patrangenaru, Vic, *Large sample theory of intrinsic and extrinsic sample means on manifolds. II*, Ann. Statist. **33** (2005), no. 3, 1225–1259.

9.1. **Working Group participants.** The following list includes everyone who registered as a participant in the working group and attended at least one session, either physically or remotely. (The SAMSI website listed 50 WG members, but a few members did not actively participate after signing up to inspect the details of the WG.) Participation ranged from fair to nearly always present.

WG members. Faculty:

1. Vic Patrangenaru (FSU) [WG leader]
2. Hongtu Zhu (UNC) [WG leader]

3. Armin Schwarzman (NCSU) [WG leader]
4. Bijan Afsari (JHU)
5. Andreas Artemiou (UK)
6. Ananda Bandulasiri (SHSU)
7. Rudy Beran (UCDavis)
8. Ted Chang (U of VA)
9. Dan Cheng (NCSU)
10. Michael Crane (FDA)
11. Josip Derado (Kennesaw State U)
12. Ian L. Dryden (U of Nottingham, UK)
13. Pang Du (VAT)
14. Leif Ellingson (TTU)
15. Giseon Heo (U of Alberta, Canada)
16. Stephan Huckemann (Göttingen, Germany)
17. Peter Kim (Guelph)
18. Yongdai Kim (Seoul National U, Sout Korea)
19. Linglong Kong (U of Alberta, Canada)
20. Michael Kostich (EPA)
21. Chirag Lakhani (HelloWallet)
22. Oleg Mahnin (UNM)
23. Bernard Omolo (USC Upstate)
24. Daniel Osborne (FAMU)
25. Robert Paige (MST)
26. Junyong Park (UMBC)
27. Marianna Pensky (UCF)
28. Nikhil Singh (UNC)
29. Valentina Staneva (JHU)
30. Samanmalee Sughatadasa (UTDallas)
31. Jiayang Sun (CWRU)
32. Nuen Tsang Yang (UCDavis)
33. Yishi Wang (UNCW)
34. Yichuan Zhao (GSU)

Postdocs:

1. Lizhen Lin (Duke) [Webmaster]
2. Wenjing Liao (Duke)

Graduate students:

1. Emil Cornea (UNC)
2. Ruite Guo (FSU)
3. David Lester (FSU)
4. Mingfei Qiu (FSU)

5. Stephen Rush (Guelph, Canada)
6. Brian St. Thomas (Duke)
7. Valentina Staneva (JHU)
8. Fabian Telschow (Göttingen, Germany)
9. Kuoadio Yao (FSU)

Topics investigated by WG.

1. Neighborhood hypotheses on Hilbert manifolds
   (led by Vic Patrangenaru)

   In this topic, motivated by a paper in JMVA by Ellingson, Patrangenaru and Ruymgaart(2013), we considered an extrinsic approach to analyzing direct similarity shapes of planar contours. We adapted Kendall's definition of direct similarity shapes of planar k-ads to shapes of planar contours, under certain regularity conditions, and Ziezold's nonparametric view on Fréchet mean shapes (Ziezold(1994)). The space of regular planar contours is dense in a Hilbert space, therefore the set of direct similarity shapes of planar contours is a subset of the complex projective space associated with this Hilbert space. This Hilbert manifold is embedded in a space of Hilbert-Schmidt operators. Using Veronese-Whitney embeddings of Kendall's Σ_2^k approximations of this embedding, for large k, one computes extrinsic sample means of direct similarity shapes of closed curves and their asymptotic distributions. To estimate these means, we appeal to a functional data analysis technique by formulating a neighborhood hypothesis testing problem on a Hilbert manifold. This methodology is applied to a one-sample test for the extrinsic mean direct similarity shape of regular contours, for which an asymptotically z-test statistic is derived. In addition, we consider using nonparametric bootstrap to approximate confidence regions for such mean shapes. Computational examples are also included for Ben Kimia's contour library. Using the contour data set, we demonstrate the computational efficiency of these techniques compared to other methods for the analysis of mean contours. For application of our technique on digital imaging data, specifically for cases when landmarks for the k-ad representation of the contours are not provided, we propose an automated randomized pseudo-landmark selection, that is useful for within population contour matching and is coherent with the underlying asymptotic theory. Note that the space of projective shapes of $3D$ configurations has a Lie group structure (Crane and Patrangenaru, 2011). Two sample tests were also considered especially in the context of projective shapes of 3D curves from digital camera images, based on the idea of Virasoro of extending the notion of Lie group to infinite dimensions.
2. Advances Shape Analysis and regression on manifolds
   (led by Hongtu Zhu and Lizhen Lin)

   Motivated by large neuroimaging data, our focus was on applications to

   - NIH normal brain development

- 1000 Functional Connectome Project
- Alzheimer's Disease Neuroimaging Initiative
- National Database for Autism Research (NDAR)
- Human Connectome Project

The methodology was aimed at manifold-valued data, under the additional assumption that the response variable is on a Riemannian manifold, that is isconnected and geodesically complete. The idea was to mimic the multivariate regression, by considering a link function on the tangent space.

Real Data Analysis (DTI) consisted in investigating early brain development by using DTI and regression models on the space of 3×3 positive definite matrices. The data consisted in DTI signals from 48 healthy infants , 18 males and 30 females, whose mean gestational age was 284 days and standard deviation 13.15 days. A 3T Allegra head only MR system was used to acquire all the images. The goal of one project was to investigate the association between the cobariates, including gender and gestational age, and DT's along the right internal capsule fiber track.

A second project was Alzheimer's Disease Neuroimaging Initiative (ADNI) The goal was to investigate the association between predictors, such as gender, age and diagnosis, and the shape of the individual's corpus callosum contour.

In addition to the intrinsic models proposed above, we are also exploring extrinsic regression models. The key ideas are to embed the manifolds of response onto some higher-dimensional Euclidean space, carry out a regression model in the Euclidean space, then project the regression estimates back onto image of the manifold. Such models are known to be computationally more efficient compared to the intrinsic models. Asymptotic theories are being developed and application are considered for a large class of manifold valued response data.

3. Nonlinear modes of variation in functional data
(led by Armin Schwartzman and Fabian Telschow)

Despite the large existing literature on functional data, few works have considered the nature of functional data objects as points on a Hilbert manifold; see (Izem and Marron, 2007), (Chen and Mueller, 2012). Since the functional data literature has focused mostly on building linear models that generalize those of multivariate analysis, that viewpoint has encountered difficulties when the variability in the data is nonlinear, such as the need for registration between curves. Often, the nonlinear variability is removed first, for example by applying a registration algorithm, and consequent analysis using linear models is performed assuming that the nonlinear variability, because it was removed, no longer exists. Ignoring the removed effects produces estimates whose standard errors are optimistic. Therefore, a full and honest analysis should include the nonlinear variability. This may be possible by the manifold viewpoint we here advocate.

To simplify the problem, the WG activity in this area began by studying the work of Izem and Marron (2007) on parametric modes of variation. In this work, rather a nonparametric registration function between curves, the nonlinear modes of variation are limited to a fixed and finite number of parameters, such as shifts or scaling of the horizontal axis, of a common template function. The approach of Izem and Marron (2007) is limited to linearly separable modes, and fails when two nonlinear modes occur simultaneously. A solution to the problem was proposed by Stephan Huckemann and Fabian Telschow to use a push-forward metric on the manifold representation of the sampled curves, so that distances on the manifold are directly mapped by means of a chart to distances between the parameter values corresponding to the nonlinear modes of variation. The main difficulty with this approach is that evaluation of the push-forward metric requires inversion of the mapping induced by the template, which is unknown. Even if it were possible to estimate the template function, a problem with interpretability remains in the sense that the distance in the parameter space depends on the relative weighting that each nonlinear mode of variation receives, which is difficult to establish since the modes are in different units.

Given the importance of template estimation, both for parametric nonlinear modes of variation and because of the need for registration in functional data, the effort of the WG in this area shifted to this topic. Under a working model that the observed curves are related by diffeomorpic warpings of the horizontal axis, the goal is to find the common template function as a Fréchet mean on an appropriate manifold representation. For the problem to be well defined, the general approach in the literature is to consider distance functions between curves that are invariant under reparametrizations; (see Tagare and Xie 2012). This is important for identifiability because any reparametrization of the template function is also a valid template function. The idea explored in the WG was to estimate the template as a Ziezold sample mean based on such an invariant distance. This is algorithmically possible. A central question however, also not answered by other methods, is whether the estimator is consistent. This remains an open question to be answered both theoretically and via simulations as work on this topic continues thanks to the initial impetus set by this WG.

WG Activities. We had a two-hour meeting every week from mid-September 2014 until the end of April 2014, with a one-month hiatus for winter break. Usually each two-hour period was split into two pieces, devoted to separate topics. Typically, one piece consisted of a presentation by a WG member, or a guest, on relevant past achievements or work in progress, while the other was active discussion on research in development by the WG, although sometimes both pieces were presentations, or the WG only discussed research.

The WG also benefitted from international visits purposefully arranged in coordination with the SAMSI program. In particular, the WG enjoyed the visit of Stephan Huckemann and Fabian Telschow from Göttingen on two occasions during the year.

Specific products in progress.

1. Circular scale space theory
   Huckemann, S. F., Kim, K.-R., Munk, A., Rehfeld, F., Sommerfeld, M., Weickert, J., Wollnik, C., *The circular SiZer, inferred persistence of shape parameters, and application to stem cell stress fibre structures* (2014). arXiv:stat.ME/1404.3300

   We generalize the SiZer of Chaudhuri and Marron (1999, 2000) for the detection of shape parameters of densities on the real line to the case of circular data. It turns out that only the wrapped Gaussian kernel gives a symmetric, strongly Lipschitz semi-group satisfying "circular" causality, i.e. not introducing possibly artificial modes with increasing levels of smoothing. Some notable differences between Euclidean and circular scale space theory are highlighted. Based on this we provide for an asymptotic theory to infer on persistence of shape features. The resulting circular mode persistence diagram is applied to the analysis of early mechanically induced differentiation in adult human stem cells from their actin- myosin filament structure. In consequence the circular SiZer based on the wrapped Gaussian kernel (WiZer) allows to discriminate at a controlled error level between three different micro-environments impacting early stem cell differentiation.
2. V. Patrangenaru, L. Ellingson, *Nonparametric statistics on manifolds and their applications*, Monographs Series in Statistics, Chapman Hall/CRC, 2014. Revision in preparation.

   The main objective of this text is to introduce the reader to a detailed and mathematically safe introduce to Object Data Analysis (ODA), and, to a limited extent, also to Big Data Analysis. While seeking answers to the fundamental question of what ODA should be all about, it is useful to go to the basic notion of variability that separates Statistics from all other sciences. One soon realizes that there are two inescapable theoretical ideas in data analysis. Firstly, one may quantify variability within or between samples only in terms of a certain distance on the sample space telling how far are observed sample points from each other. Secondly, the distance, as a function of the two data points separated by it, has to have some continuity property, to make any consistency statement possible justifying why the larger the sample, the closer the sample variability measure to its population counterpart. In addition, since an asymptotic theory based on random observations is necessary to estimate the population variance based on a large sample, such a theory can be formulated only under the additional assumption of differentiability of some power of the square distance function. In summary, ODA imposes some sort of differentiable structure on the sample space that has to be consequently either a manifold, or having some manifold related structure, no matter what the nature of the objects is.

   However the overwhelming number of Statistics users are specializing more in understanding the nature of the objects themselves, having little or no exposure to the basics of geometry and topology of manifolds knowledge needed to develop appropriate of methodology for ODA. At the same time, theoretical mathematicians who have a reasonable knowledge about manifolds might be unfamiliar with nonparametric multivariate statistics, while computational grad students and computational data analysts involved with Object Data sometimes ask for a sound nonparametric statistics explanation, or for a brief multidimensional differential geometry or topology toolkit, that may help them design fast algorithms for ODA. To answer such demands, we structured our monograph as follows. We first introduce the basics for the three "pillars of ODA": (i) examples of object data, (ii) nonparametric multivariate statistics and (iii) geometry and topology of manifolds. Secondly we develop a general methodology based on (i) and (ii), and "translate" this methodology, in the context of certain manifolds arising in statistics. Finally we apply this methodology to concrete examples of ODA.

3. E. Osborne, V. Patrangenaru, M. Qiu, (2014). Shape analysis methods in medical imaging. *Festschrift volume for Kanti Mardia*, Ed. J. Kent and I.L. Dryden, in preparation.
4. R. Paige, V. Patrangenaru, M. Qiu, (2014). Statistical analysis of projective shapes of 3D curves, in preparation.
5. J. Derado and V. Patrangenaru, (2014). Virasoro like Algebras and two sample tests for mean projective shapes of 3D curves, in preparation.
6. R. Guo, V. Patrangenaru and K.D. Yao, (2014). Cartan means and Cartan antimeans on stratified spaces (Invited presentation at INSPS II-Cadiz, Spain), in preparation.

Publications directly resulting from WG research.

Papers submitted/accepted/published in refereed journals:

1. R. N. Bhattacharya and V. Patrangenaru, *Statistics on manifolds and landmarks based image analysis: a nonparametric theory with applications*, Journal of Statistical Planning and Inference **145** (2014), 1–22.

   This article provides an rejoinder for the discussions of our paper on recent developments in nonparametric inference on manifolds, along with a brief account of an emerging theory on data analysis on stratified spaces.

Papers submitted/accepted/published in other refereed venues

1. Semi-intrinsic statistical analysis Huckemann, S. F., *(Semi-)intrinsic statistical analysis on non-euclidean spaces*, In Advances in Complex Data Modeling and Computational Methods in Statistics, Springer (2014), to appear.

   Often, applications from biology and medical imaging lead to data on non-Euclidean spaces. On such spaces the Euclidean concept of a mean forks into several canonical generalizations of non-Euclidean means. More involved data descriptors, for instance principal components generalize into even more complicated concepts. (Semi)-intrinsic statistical analysis allows to study inference on descriptors that can be represented as elements of another non-Euclidean space. We give examples for geodesic principal components on shape spaces, concentric small circles on spheres and configurations on rotation groups. In particular, with respect to the statistical inference via central limit theorems, due to the geometry of the spaces, there are curious non-Euclidean phenomena.

2. Qiu, V. Patrangenaru, L. Ellingson, *How far is the corpus callosum of an average individual from Albert Einstein's?*, Proceedings of COMPSTAT 2014, Geneva.

   The optic nerves meet at the Optic Chiasma (OC) in a midsagittal plane at the base of the brain. There, half of the axons from each nerve cross over into the other nerve, so that some visual information from the left eye travels in parallel with information from the right eye within each of the two nerves. The blending of the two eye images allows one to perceive the projective shape of the scene. The Corpus Callosum (CC) connects the two cerebral hemispheres and facilitates interhemispheric communication. It is the largest white matter structure in the brain. Albert Einstein's brain was removed shortly after his death, weighted, dissected and photographed by a pathologist. High resolution versions of those pictures were quantitatively studied in two recent papers listed in the references. Contours of CC midsagittal sections are extracted from MRI images. Given that Einstein passed at 76, we extracted a small subsample of CC brain contour, in the age group 64-83, and tested how far is the average CC contour from Einstein's. The analysis was performed on the Hilbert manifold of planar contours, following the methodology recently developed by the authors.

Other publications related to work done at SAMSI.
Papers submitted/accepted/published in refereed journals:

1. Chen, Y., Du, P., and Wang, Y., "Variable selection in linear models", WIREs *Computational Statistics* **6** (2014), 1–9.
2. Du, P. and Wang, X., "Penalized likelihood functional regression", *Statistica Sinica* (2014), accepted.
3. L. Ellingson, D. Groisser, D. Osborne, V. Patrangenaru and A. Schwartzman, *Nonparametric bootstrap of sample means of positive-definite matrices with an application to Diffusion Tensor Imaging data analysis*, Computational Statistics & Data Analysis (2014), submitted.
4. Kachouie N, Gerke T, Winter J, Huybers P, Schwartzman A. (2014), Nonparametric regression for estimation of spatial and temporal mountain glacier retreat from satellite images. *submitted*
5. Cornea, E., Zhu, H.T., and Ibrahim, J. G. (2014). Intrinsic regression model for data in Riemannian symmetric space. *JRSS, Series B, under revision.*
6. Huang, C., Styner, M., and Zhu, H.T. (2014). Penalized mixtures of offset-normal shape factor analyzers with application in clustering high-dimensional shape data. *Journal of American Statistical Association*, under revision.

7. Zhu, H.T., Khondker, Z. S., Lu, Z.H., and Ibrahim, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, in press.
8. Sun, Q., Zhu, H.T., Liu, Y. F., and Ibrahim, J.G. (2014). SPReM: Sparse Projection Regression Model for High-dimensional Linear Regression. *Journal of the American Statistical Association*, in press.
9. Mihye, A., Shen, H.P., Lin, W. L., and Zhu, H.T., (2014). A sparse reduced rank framework for group analysis of functional neuroimaging data. *Statistica Sinica*, in press.
10. Gao, Q. B., Mihye, A. and Zhu, H.T., (2014). Cook's distance measures for varying coefficient models with functional response. *Technometrics*, in press.
11. Guo, R.X., Ahn Mihye, and Zhu, HT., (2014). Spatially weighted principal component analysis for imaging classification, *Journal of Computational and Graphical Statistics*, in press.
12. Zhu, H.T., Fan, J.Q., and Kong, L.L., (2014). Spatially varying coefficient models with applications in neuroimaging data with jumping discontinuity. *Journal of the American Statistical Association*, in press.
13. J. W. Hyun, Li, Y. M., Gilmore, J., Lu, Z.H., Styner, M., and Zhu, H.T. (2014). SGPP: Spatial Gaussian Predictive Process Models for Neuroimaging Data. *NeuroImage*, **89**, 70–80.
14. Ford AL, An H, Kong L, Zhu H, Vo KD, Powers WJ, Lin W, and Lee JM (2014). Clinically-relevant reperfusion in acute ischemic stroke: MTT performs better than Tmax and TTP. *Translational Stroke Research*, in press.
15. A. R. Verde, F. Budin, J.B. Berger, A. Gupta, M.Farzinfar, A. Kaiser, M. Ahn, H.J Johnson, J. Matsui, H.C. Hazlett, A. Sharma, C. Goodlett, Y. Shi, S. Gouttard, C. Vachet, J. Piven, H. Zhu, G. Gerig, M. A.Styner., (2013). UNC-Utah NA-MIC Framework for DTI Fiber Tract Analysis, Frontiers in Neuroinformatics.
16. Yuan, Y., Gilmore, J., Geng, X. J., Styner, M., Chen, K. H., Wang, J. L., and Zhu, H.T., (2013). A longitudinal functional analysis framework for analysis of white matter tract statistics. *NeuroImage*, in press.
17. V. Patrangenaru, M. Qiu and M. Buibas, (2014). Two Sample Tests for Mean 3D Projective Shapes from Digital Camera Images. *Methodology and Computing in Applied Probability* **16**, 485–506.

Grant proposals.

1. *Sy-Miin Chow (Penn State, Human Development and Family Studies)
   NSF SES 1357666
   9/15/2014–8/31/2017
   Developing dynamic tools for analyzing irregularly spaced longitudinal affect data
   Role: Co-Principal Investigator
   Total Direct Cost: $350,001

2. Hongtu Zhu
   NSF DMS
   9/15/2014–8/31/2017
   Advanced Statistical Methods for Functional Imaging Data
   Role: Principal Investigator
3. Vic Patrangenaru
   NSA MSP
   5/01/2014–4/30/2016
   Nonparametric statistical analysis of spatial scenes from digital camera images
   Role: Principal Investigator
4. Pang Du
   NSF MSP
   Collaborative Research: Robust and flexible analysis of big data in bioinformatics – Feature selection, graphical models, and structure determination
   Role: Principal Investigator

Conference and workshop presentations.

1. Vic Patrangenaru (including contributions by many of our WG members), LDHD Transition Workshop
2. Emil Cornea (including contributions by Hongtu Zhu), LDHD Transition Workshop
3. Vic Patrangenaru, SAMSI–CANSSI workshop "Geometric Topological and Graphical Model Methods in Statistics" at Fields Institute, Toronto, Canada, May 22–23, 2014
4. Stephan Huckemann, SAMSI–CANSSI workshop on "Geometric Topological and Graphical Model Methods in Statistics" at Fields Institute, Toronto, Canada, May 22–23, 2014
5. At least 5 posters have been presented by group members at these two LDHD-related workshops
6. Huckemann, Paige, and Patrangenaru will attend the second conference of the International Society of Nonparametric Statistics, June 11–17, Cadiz, Spain, in a session of "Data analysis on stratified spaces"
7. Stephan Huckemann, COMPSTAT 2014, Image Analysis session, August 19–22, 2014
8. Vic Patrangenaru, COMPSTAT 2014, Image Analysis session, August 19–22, 2014

10. Online streaming and sketching

The focus was on methodology and fast algorithms for computing leverage scores, with application to astronomy and genomics. Specific topics included:

- Leverage scores: computation, fast approximation, sensitivity, numerical stability of algorithms, behavior under sketching
- Randomized low-rank approximations: subset selection, CUR, Nystrm, PCA, robust PCA, subsampled regression, regression on manifolds, construction of robust linear models, windowed and online streaming approaches
- Randomized sketching and importance sampling strategies: methodology and numerical computation
- Local vs global: eigenvector and invariant subspace localization, eigen-analysis of data connectivity matrices, numerical stability of streaming and updating methods, relation to generalized eigenvalue problems
- Data fusion/integration: robust and fast/streaming methods with application to galaxy formation and evolution

WG members. Active faculty:

1. Tamás Budavári, Johns Hopkins University (Physics & Astro)
2. *Ilse Ipsen, North Carolina State University (Math) [WG leader]
3. Michael Mahoney, Stanford University (Math) [WG leader]

Postdocs:

1. David Lawlor, SAMSI and Duke University (Math) [WG leader]

Graduate students:

1. Armin Eftekhari, Colorado School of Mines (EE and CS)
2. John Holodnak, North Carolina State University (Math)
3. Thomas Wentworth, North Carolina State University (Math)

Occasional members:

1. Andreas Artemiou, Cardiff University, UK (Math)
2. Haim Avron, IBM
3. Christos Boutsidis, Yahoo! Inc.
4. Petros Drineas, Rensselaer Polytechnic Institute (CS)
5. *Wenjing Liao, SAMSI postdoc
6. Minh Pham, SAMSI postdoc
7. Johan Van Horebeek, CIMAT, Mexico (CS)

Topics investigated by WG.

1. *Randomized approximation of the Gram matrix: exact computation and probabilistic bounds* (John Holodnak, Ilse Ipsen)

   Given a real matrix A with n columns, the problem is to approximate the Gram product AA^T by $c \ll n$ weighted outer products of columns of A. Necessary and sufficient conditions for the exact computation of AA^T (in exact arithmetic) from $c \geq \text{rank}(A)$ columns depend on the right singular vector matrix of A.

   For a Monte-Carlo matrix multiplication algorithm by Drineas et al. that samples outer products, we present probabilistic bounds for the 2-norm relative error due to randomization. The bounds depend on the stable rank or the rank of A, but not on the matrix dimensions. Numerical experiments illustrate that the bounds are informative, even for stringent success probabilities and matrices of small dimension. We also derive bounds for the smallest singular value and the condition number of matrices obtained by sampling rows from orthonormal matrices.

2. *The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems* (Ilse Ipsen, Thomas Wentworth)

   Motivated by the least squares solver *Blendenpik*, we investigate three strategies for uniform sampling of rows from $m \times n$ matrices Q with orthonormal columns. The goal is to determine, with high probability, how many rows are required so that the sampled matrices have full rank and are well-conditioned with respect to inversion.

   Extensive numerical experiments illustrate that the three sampling strategies (without replacement, with replacement, and Bernoulli sampling) behave almost identically, for small to moderate amounts of sampling. In particular, sampled matrices of full rank tend to have two-norm condition numbers of at most 10.

   We derive a bound on the condition number of the sampled matrices in terms of the coherence μ of Q. This bound applies to all three different sampling strategies; it implies a, not necessarily tight, lower bound of $\mathcal{O}(m\mu \ln n)$ for the number of sampled rows; and it is realistic and informative even for matrices of small dimension and the stringent requirement of a 99% success probability.

   For uniform sampling with replacement we derive a potentially tighter condition number bound in terms of the leverage scores of Q. To obtain a more easily computable version of this bound, in terms of just the largest leverage scores, we first derive a general bound on the two-norm of diagonally scaled matrices.

   To facilitate the numerical experiments and test the tightness of the bounds, we present algorithms to generate matrices with user-specified coherence and leverage scores. These algorithms, the three sampling strategies, and a large variety of condition number bounds are implemented in the Matlab toolbox *kappa_SQ_v3*.

3. *Sensitivity of Leverage Scores* (Ilse Ipsen, Thomas Wentworth)

The sampling strategies in many randomized matrix algorithms are, either explicitly or implicitly, controlled by statistical quantities called leverage scores. We present four bounds for the sensitivity of leverage scores as well as an upper bound for the principal angles between two matrices. These bounds are expressed by considering two real $m \times n$ matrices of full column rank, A and B. Our bounds shows that if the principal angles between A and B are small, then the leverage scores of B are close to the leverage scores of A. Next, we show that the principal angles can be bounded above by the two-norm condition number of A, $\kappa(A)$ and $\|B - A\|_2$. Finally, we combine these bounds and derive bounds for the leverage scores of B in terms of $\kappa(A)$ and $\|B - A\|_2/\|A\|_2$ and show that if $\|B - A\|_2/\|A\|_2$ and $\kappa(A)$ are small, then the leverage scores of B are close to the leverage scores of A.

4. *Kappa_SQ: A Matlab Package for Randomized Sampling of Matrices with Orthonormal Columns* (Ilse Ipsen, Thomas Wentworth)

   The kappa_SQ software package is designed to assist researchers working on randomized row sampling. The package contains a collection of `Matlab` functions along with a GUI that ties them all together and provides a platform for the user to perform experiments.

   In particular, kappa_SQ is designed to do experiments related to the two-norm condition number of a sampled matrix, $\kappa(SQ)$, where S is a row sampling matrix and Q is a tall and skinny matrix with orthonormal columns. Via a simple GUI, kappa_SQ can generate test matrices, perform various types of row sampling, measure $\kappa(SQ)$, calculate bounds and produce high quality plots of the results. All of the important codes are written in separate `Matlab` function files in a standard format which makes it easy for a user to either use the codes by themselves or incorporate their own codes into the kappa_SQ package.

5. *Global and local connectivity analysis of galactic spectra* (David Lawlor, Tamás Budavári, Michael Mahoney)

   In the past decade much attention has been paid to examining the connectivity of data sets as a means to extract low-dimensional structure from nominally high-dimensional data. This is usually accomplished by computing the leading eigenvectors of the Laplacian of the data connectivity graph, which are inherently global quantities taking into account the interactions among all data points. In data sets containing multiple subpopulations this may be disadvantageous, and a more local approach may be appropriate.

   We investigate theoretical and empirical properties of such an approach, the recent semi-supervised eigenvectors of Mahoney et al. This project explores the use of semi-supervised eigenvectors for classifying galactic spectra from the Sloan Digital Sky Survey. We also present a thorough empirical investigation of the properties of embeddings via global eigenvectors of graph Laplacians.

6. *Semi-supervised eigenvectors and constrained eigenvalue problems* (David Lawlor, Ilse Ipsen)

We recast the optimization framework in which the semi-supervised eigenvectors of Mahoney et al. are defined as a nonlinearly constrained eigenvalue problem. We investigate methods for the solution of this type of problem in a general setting for both sparse and dense matrices.

7. *Regression in high dimension via geometric multi-resolution analysis* (David Lawlor and Mauro Maggioni)

   We present a framework for high-dimensional regression using the GMRA data structure. In analogy to a classical wavelet decomposition of function spaces, a GMRA is a tree-based decomposition of a data set into local linear projections. Moreover, for new points, GMRA admits a fast algorithm for computing the projection coefficients on the already-learned dictionary. Within each node of the tree one can also assign regression coefficients in any manner; here we study the simple case of weighted linear regression. Empirical results show improvements in both accuracy and computational speed on both synthetic and real data sets. We prove theorems guaranteeing rates of approximation given some (mild) smoothness assumptions on both the manifold and the regression function.

Specific products in progress.

1. *Sensitivity of Leverage Scores* (Ilse Ipsen, Thomas Wentworth), to be submitted for publication. arXiv:math.NA/1402.0957
2. *Kappa_SQ: A Matlab Package for Randomized Sampling of Matrices with Orthonormal Columns* (Ilse Ipsen, Thomas Wentworth), to be submitted for publication. arXiv:math.NA/1402.0642
3. *Global and local connectivity analysis of galactic spectra* (David Lawlor, Tamás Budavári, Michael Mahoney), in preparation.
4. *Regression in high dimension via geometric multi-resolution analysis* (David Lawlor and Mauro Maggioni), in preparation.
5. *Semi-supervised eigenvectors and constrained eigenvalue problems* (David Lawlor, Ilse Ipsen), in preparation.

Publications directly resulting from WG research.

Papers submitted/accepted/published in refereed journals:

1. John Holodnak, Ilse Ipsen, *Randomized Approximation of the Gram Matrix: Exact Computation and Probabilistic Bounds*, SIAM J. Matrix Anal. Appl., under revision. arXiv:math.DS/1310.1052
2. Ilse Ipsen, Thomas Wentworth, *The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems*, SIAM J. Matrix Anal. Appl., under revision.
   arXiv:math.NA/1203.4809

Publicly available software.

1. Ilse Ipsen, Thomas Wentworth, *Kappa_SQ*. arXiv:math.NA/1402.0642

Organization of follow-up and related workshops.

1. David Lawlor, Garvesh Raskutti, *Randomized algorithms for statistical inference*, International Symposium on Business and Industrial Statistics/Conference of ASA Section on Statistical Learning and Data Mining, Durham, NC, 9–11 June 2014. http://www.stat.duke.edu/~banks/dcc

Conference and workshop presentations.

1. Ilse Ipsen, *Introduction to randomized matrix algorithms*, AMS Southeastern Spring Sectional Meeting, University of Knoxville, TN, 21–23 March 2014. http://www.ams.org/meetings/sectional/2216_special.html
2. Ilse Ipsen, *Rolling the Dice on Big Data*, SAMSI LDHD Education & Outreach Undergraduate Workshop, 20 February 2014.
3. John Holodnak, *Sensitvity of Leverage Scores to Perturbations* (poster), LDHD Transition Workshop, 12–14 May 2014.
4. Ilse Ipsen, *A (subjective) Introduction to Randomized Matrix Algorithms*, International Workshop on Accurate Solution of Eigenvalue Problems X, Drubovnik, Croatia, 2–5 June 2014. http://iwasep.fesb.hr/iwasep10
5. Ilse Ipsen, *Randomized Algorithms for Numerical Linear Algebra*, Householder Symposium XIX, Spa, Belgium, 8–13 June 2014. http://sites.uclouvain.be/HHXIX
6. Ilse Ipsen *Leverage scores: Sensitivity and an App*, Workshop on Algorithms for Modern Massive Datasets, UC Berkeley, 17–20 June 2014. http://mmds-data.org
7. John Holodnak, *Sensitvity of Leverage Scores to Perturbations*, SIAM Annual Meeting, Chicago, 7–11 July 2014. http://www.siam.org/meetings/an14
8. Ilse Ipsen, *Probabilistic Bounds for a Randomized Preconditioner for a Krylov Least Squares Solver*, SIAM Annual Meeting, Chicago, 7–11 July 2014. http://www.siam.org/meetings/an14
9. Ilse Ipsen, *An Introduction to Randomized Matrix Algorithms*, 5th International Conference on Numerical Linear Algebra and Scientific Computing, Shanghai, 24–30 October 2014. http://lsec.cc.ac.cn/~NASCNAG
10. David Lawlor, *Regression in high dimensions via geometric multi-resolution analysis* (poster), LDHD: Statistical Inference in Sparse High-Dimensional Models, 24–26 February 2014.
11. David Lawlor, *Regression in high dimensions via geometric multi-resolution analysis* (poster), LDHD: SAMSI-CRM Workshop on Geometric Aspects of High-Dimensional Inference, 31 March – 2 April 2014.
12. David Lawlor, *Global and local connectivity analysis of galactic spectra*, LDHD Transition Workshop, 12–14 May 2014.
13. David Lawlor, *Regression in high dimensions via geometric multi-resolution analysis* (poster), International Symp. on Business and Industrial Statistics/Conference of the ASA Section on Statistical Learning and Data Mining, Durham, NC, 9–11 June 2014. http://www.stat.duke.edu/~banks/dcc

14. David Lawlor, *Global and local connectivity analysis of galactic spectra* (poster), Workshop on Algorithms for Modern Massive Datasets, UC Berkeley, 17–20 June 2014. http://mmds-data.org
15. David Lawlor, *Regression in high dimensions via geometric multi-resolution analysis* (poster), SIAM Annual Meeting, Chicago, 7–11 July 2014. http://siam.org/meetings/an14

11. Graduate students

Under the leadership of Max Sommerfield and Hamza Ghadyali, a weekly meeting for SAMSI graduate students was organized in the Fall of 2013 at SAMSI. In the two to three hour long weekly meetings, students gave both informal talks and formal presentations of their research. In the informal talks, participants asked very detailed questions to fully understand the problem which then resulted in much creative brainstorming during the meetings. This informal structure differed from any usual seminar talks in which many participants often either do not ask questions to understand the problem, or even when they understand the problem, they may not freely contribute their ideas. A partial list of the talks given is listed below, followed by some comments from the participants on how this working group benefitted them personally.

WG members. Graduate students:

1. Hamza Ghadyali
2. Max Sommerfield
3. Vahan Huroyan
4. John Goes
5. Stephen Rush
6. Thomas Wentworth
7. Fabian Telschow
8. Polat Charyyev
9. Nanwei Wang
10. Chong Shao

WG Activities.

1. Vahan Huroyan: Distributed robust PCA parallel eigenvalue decomposition
2. Thomas Wentworth: Generating random matrices with prescribed leverage scores
3. Nanwei Wang: Maximum likelihood estimation and geometric properties in discrete graphical models
4. Nanwei Wang: Kronecker graphs: an approach to modeling networks high dimensional covariance matrix estimation by graphical lasso
5. John Goes: Stochastic formulations of robust subspace recovery natural extensions of highdimensional PCA eigenvalue distribution of random matrices
6. Polat Charyyev: Optimal obstacle placement problem with disambiguations
7. Chong Shao: Geodesic PCA
8. Hamza Ghadyali: ECG-derived respiratory signals
9. Hamza Ghadyali: Probabilistic Fréchet means of persistence diagrams
10. Max Sommerfield: Circular SiZer with an application to early stem cell differentiation
11. Max Sommerfield: Algorithmic aspects of support estimation
12. Max Sommerfield: uncertainty in excursion sets

Participant feedback.
"I greatly benefitted from the variety of multicultural interaction and diversity of thoughts, which made serious difference both in my research and personal life."
–Nanwei Wang

"The informal nature of the group made working together to solve research problems very productive. I got many good ideas for my own research and I very much enjoyed trying to help other's with their research problems. We were also able to work on the problems more in depth that in the other working groups."
–Thomas Wentworth

"... the meetings were a nice way to know each others research interests and know some new directions in the field."
–Vahan Huroyan