Final Report for the SAMSI Program: Analysis of Object Oriented Data (2010 - 2011)

Background

Modern science is generating a need to understand, and statistically analyze, populations of increasingly complex types. The term "Analysis of Object Oriented Data" (AOOD) is aimed at encompassing a broad array of such methods. This SAMSI program brought together a diverse group of researchers (from statistics, other parts of mathematics, and related sciences) to explore the common structure that underlies such methodologies, and to use this knowledge to motivate and synthesize new approaches.

This was a year-long SAMSI program for 2010-2011 on the analysis of complex data types that are an extension of Functional Data Analysis where one considers methods to analyze data samples of complex objects.

Organizers: Hans-Georg Müller (University of California, Davis), Jane-Ling Wang (University of California, Davis) (Co-Leaders); Ian Dryden (University of South Carolina), Steve Marron (University of North Carolina), Jim Ramsay (McGill University).

Local Scientific Coordinator: Steve Marron (UNC).

Directorate Liaison: James Berger (SAMSI); Richard Smith (after 7/1/10)

National Advisory Committee Liaison: Jiangqing Fan (Princeton).

1 Workshops

1.1 Opening Workshop September 12-15, 2010

1.1.1 Summary

The Opening Workshop for the SAMSI program on Analysis of Object Data (AOOD) was held on Sunday-Wednesday, September 12-15, 2010, at the Radisson RTP in Research Triangle Park, NC.

Tutorial sessions for each of five threads took place on Sunday, September 12. Invited talks were presented Monday to Wednesday. There was a poster session and reception on Monday, September 13. Immediately following the workshop, on Thursday and Friday, research working groups convened for initial meetings at SAMSI.

1.1.2 Activities

The workshop focused on five threads that exemplify the AOOD idea of generalizing the functional data analysis concept of random curves as data points to more general objects as data points. These include objects that are Euclidean, i.e., (constant length) vectors of real numbers, mildly non-Euclidean, i.e. points on a manifold and shapes, or strongly non-Euclidean, i.e. tree or graph structured objects. The five focus areas of the workshop were

1. Functional Data Analysis, theory and applications for samples of curve and surface data in the life, social, environmental and physical sciences and the interface with longitudinal data analysis;

2. Time Dynamics Data, with an emphasis on methodology for the analysis of observations that are governed by differential equations and the modeling of the dynamics of growth, gene expression, infections or online auctions;

3. Shape Analysis and Manifold Data, including the analysis of landmarks, curves, surfaces, and volumes and the analysis of data that naturally lie on a manifold such as directional data;

4. Image Analysis, with a focus on applications where data consist of a sample of images, including fMRI data, diffusion tensor imaging and diffeomorphisms for brain mapping;

5. Tree and Graph Structured Data, which are "strongly non-Euclidean" and require the development of statistical analysis from the ground up, to analyze samples of trees and graphs.

The workshop aimed at emphasizing synergies and interactions between these threads, and culminated in the formation of research working groups in the afternoon of Wednesday, September 15. These working groups then met individually at SAMSI on Thursday and Friday to further address specific research objectives to be addressed by the working group over the ensuing year. These meetings also established modes of cooperation for the working groups, via web or teleconference, to facilitate full participation of all members, regardless of residence status at SAMSI.

The workshop was full to capacity with about 150 participants. On the first day five tutorial lectures were held to provide extensive background of the five themes of the program. In particular the tutorials were:

- FDA Tutorial: Functional Data Analysis and Related Topics, Fang Yao, University of Toronto
- Dynamics Tutorial: Models for Output-Buffered Systems: An Introduction to Dynamics, Jim Ramsay, McGill University
- Images Tutorial: Multivariate Statistical Analysis of Deformation Momenta Relating Anatomical Shape to Neuropsychology, Sarang Joshi, University of Utah and Martin Lindquist, Columbia University
- Shapes and Manifolds Tutorial: Shape Analysis and Manifold Data, John Kent, University of Leeds
- Trees Tutorial: Trees as Data, J.S. Marron, University of North Carolina

Videos of the expository lectures are available on the SAMSI website.

There then followed five half-day sessions on each of the themes. Each session consisted of lectures on future challenges from distinguished experts and also excellent contributions from new researchers.

- **Functional Data Analysis**

  Chair: Hans-Georg Müller, University of Calfornia, Davis.

  - Alois Kneip, University of Bonn, "Challenges for FDA Theory and Methodology."
  - Naisyin Wang, University of Michigan, "Challenges for FDA in Longitudinal Studies."
  - Shuang Wu, University of Rochester, " FDA for Non-traditional Data."
  - Ciprian Crainiceanu, John Hopkins University, " My first 100 Terabytes of Data: Challenges for FDA Modeling."

- **Dynamics**

  Chair: Jim Ramsay, McGill University.

  - Aaron King, University of Michigan, " Plug-and-play Inference for Stochastic Dynamical Systems."
  - Sy-Miin Chow, University of North Carolina, " Applications of Differential Equation Modeling in the Social Sciences."

- **Images**

  Chair: Jane-Ling Wang, University of California-Davis.

  - Dubois Bowman, Emory University, "Statistical Modeling of Brain Imaging Data: An Overview, Challenges, and Future Directions."
  - Laurent Younes, Johns Hopkins University, "Shape Analysis of Diastolic/Systolic Paired Cardiac Images."
  - Armin Schwartzman, Harvard School of Public Health, " Data Objects in Diffusion Tensor Imaging."
  - Jonathan Taylor, Stanford University, " Overview of Statistical Inference, Multiple Comparisons and Gaussian Random Fields for Neuroimaging Data."

- **Trees**

  Chair: J.S. Marron, University of North Carolina - Chapel Hill.

  - Burcu Aydin, Hewlett Packard Research, "Principal Component Analyses for Trees."
  - Sean Skwerer, University of North Carolina, "Analysis of Object Data: Averaging Metric Trees."
  - Shanker Bhamidi, University of North Carolina, " Dyck Path Correspondence and the Statistical Analysis of Brain Vascular Networks."
  - Haonan Wang, Colorado State University, "Smoothing and Branching Process Inference on Trees."

- **Shapes and Manifolds**

  Chair: Ian Dryden, University of South Carolina.

  - Huiling Le, University of Nottingham, " Some Aspects of Statistics on Riemannian Manifolds from the Perspective of Shape Analysis."
  - Anuj Srivastava, Florida State University, " Towards Statistical Modeling of Shapes of Curves and Surfaces."
  - Victor Panaretos, EPFL Lausanne, " Statistical Shape and Random Tomography in Structural Biology."
  - Alain Trouvé, Ecole Normale Supérieure de Cachan, " From Shape Comparison to Shape Evolution with a Geometrical and Statistical Perspective."

During the workshop there were a couple of Minute Madness sessions, where a large number of the participants had two minutes to present a topic of interest with a very strict time limit and a couple of slides. These lively sessions gave a showcase to the broad range of research interests of the participants. In addition, there was a poster session and reception, with a large number of excellent posters.

A key aspect of the opening workshop was the formation of the working groups, which was critical to the subsequent developments of the program.

The AOOD program was privileged to have such an outstanding workshop to set the scene for the year long research activities. Each theme's activities included presentations from some of the leading researchers in each area, from outstanding new researchers and involvement and interaction with a very knowledgeable and expert audience. The chairs of each session ensured that there was a lively amount of discussion throughout, and the workshop gave the perfect start to the program.

1.2 Workshop on the Interface Functional and Longitudinal Data Analysis - November 8-10, 2010 at SAMSI

1.2.1 General Description

Longitudinal and time-dynamic data are collected throughout the life sciences, social sciences and environmental sciences to explore time-dynamic aspects of phenomena that evolve over time. Longitudinal data analysis is a core technique of biostatistics, and has seen fast development over the last decades, primarily in the area of random effects modeling. Several researchers have begun more recently to explore the potential of adopting a functional perspective for the analysis of longitudinal data. The promise is that Functional Data Analysis provides flexible approaches dependence structures, includes derivatives and the analysis of dynamics, and permits the data to determine the most suitable model.

The workshop integrated these various perspectives and participants identified promising technology and prospective areas of interest for future research along this interface, emphasizing both methodology and applications. The workshop brought together researchers working in separate communities and explored and harnessed synergies between the various approaches. The workshop succeeded in generating lively discussions and engagement by all participants. Throughout these discussions, several key issues such as the role of functional principal components and the signficance of functional data analysis in biomedical longitudinal studies were thoroughly debated.

Organizers: Marie Davidian (N.C. State University), Fang Yao (University of Toronto), Hans-Georg Müller (Univ. of California-Davis)

Topics:

- Multivariate longitudinal/functional data
- Multilevel longitudinal/functional data
- Functional components in joint modeling of longitudinal and survival outcomes
- Longitudinal/functional data as components in regression models
- Longitudinal/functional models for dynamics
- Variable selection in longitudinal/functional models

1.2.2 Talks and Events

1. Ray Carroll, Texas A&M University: Generalized Functional Linear Models with Semiparametric Single-Index Interactions
2. Damla Şentürk, Penn State University: Functional Varying Coefficient Models for Longitudinal Data
3. Wensheng Guo, University of Pennsylvania: Functional Mixed Effects Spectral Models
4. Discussion and Connections to Working Groups
5. Tailen Hsing, University of Michigan: Uniform Convergence Rates for Principal Component Analysis in Functional/ Longitudinal Data
6. Jeff Morris, University of Texas: Adaptive, Robust Functional and Image Regression in Functional Mixed Models
7. Fang Yao, University of Toronto: Additive Modeling of Functional Regression and its Gradients
8. Discussion and Connections to Working Groups
9. Poster Session and Reception
10. Xihong Lin, Harvard University: Likelihood Based and Estimating Equation Based Methods for Variable Selection
11. Helen Zhang, North Carolina State University: Variance Component Selection in Linear Mixed Models
12. Graciela Boente, Universidad de Buenos Aires and CONICET: Robust Functional Principal Components: a projection-pursuit approach
13. Discussion and Connections to Working Groups
14. Jim Ramsay, McGill University: Economical Models for Functional Covariation
15. Jie Peng, University of California-Davis: Fitting Ordinary Differential Equation Models with Longitudinal Data
16. Huilin Wu, University of Rochester: Comparing Functional Data Analysis Approach and Nonparametric Mixed-Effects Modeling Approach for Longitudinal Data Analysis
17. Discussion and Connections to Working Groups
18. Ana-Maria Staicu, N.C. State University: Skewed Functional Processes and their Applications
19. Jimin Ding, Washington University: Time-varying Coefficient Cox model with Nonparametric Longitudinal Covariates
20. Discussion and Connections to Working Groups

1.3 Workshop on the AOOD Meets Evolutionary Biology – April 30-May 2, 2011 at SAMSI

1.3.1 General Description

The goal of this workshop was to build a bridge between activities in the SAMSI Analysis of Object Data Program, and the area of Function Valued Traits in Evolutionary Biology. The latter field already has strong ties with Functional Data Analysis, and this workshop was intended to define the new area of Object Valued Traits in Evolutionary Biology. Specific data contexts where the object valued viewpoint was specifically aimed included fly wing shapes (intersecting with the Shape and Manifold Research in AOOD), caterpillar growth trajectories (intersecting with Time Dynamics in AOOD) and rat brain blood vessel trees, plus families of viral phylogenetic trees (intersecting with the Tree Research in AOOD).

1.3.2 Highlights

Highlights of the meeting were:

- An Opening 2-Minute Madness Session, where each participant gave a 2 minute introductory talk. This enabled this diverse group of people to understand the breadth of attendees, and also to find people they had not previously known, to engage in individual conversations.
- A series of talks by biological leaders in Function Valued Traits, aimed at generating discussion among OODA researchers:
  - Joel Kingsolver, University of North Carolina, Genetic Variation and Evolution of Function-Valued Traits
  - Jay Beder, Univ. of Wisconsin-Milwaukee, Estimating the selection gradient of a function-valued trait
  - David Houle, Florida State University, Connecting Phenomic Objects to Genomic Predictors
  - Patrick Carter, Washington State University, Evolution of the Integrated Phenotype: A Function Valued Approach
  - Nancy Heckman, University of British Columbia, Dependence in Functional Data Analysis
  - Heather Jamniczky, University of Calgary, Quantification of Unusual Biological Shapes in Three Dimensions
  - Washington Mio, Florida State University, Spectral Methods in Shape Analysis
  - Saunak Sen, Univ. of California-San Francisco, Genetic mapping of function-valued traits
- A series of talks by OODA members on recently developed ideas related to evolutionary biology:
  - Daniel Gervini, University of Wisconsin, Semiparametric Curve Registration
  - Sarang Joshi, University of Utah, Towards Imaging Based Biomarkers
  - John Aston, Warwick University, A step towards a function-valued typology for language
  - J.S. Marron, University of North Carolina, OODA of Tree Structured Objects
  - Ezra Miller, Duke University, Stratified statistics for evolutionary biology

- At various points there were discussions at a number of levels, with excellent interactions between the two main groups.

1.4 Transition Workshop, June 9-11, 2011

1.4.1 Summary

The goals for this workshop were to review and discuss the progress made during the year on the various program projects and their synergies. Current status and future directions of research in the program were assessed. The workshop featured sessions on the five thematic areas of this program: functional data analysis, dynamics, data on manifolds, brain imaging and trees; and it brought together participants from the various working groups in these areas.

The organized discussions during the Workshop and the informal discussions after paper presentations revealed many important connections between these five themes. It was recognized that curve, surface and volume registration problems in functional data analysis and brain imaging were essentially shape analysis problems under some specific constraints, and a follow-up workshop at the next Joint Statistical Meetings was discussed as an outcome. The possibility of representing many functional and spatial data analysis problems as dynamic systems through ordinary and partial differential equations was frequently commented upon, and considerable progress along these lines during the year was outlined.

Although the phrase "strongly non-Euclidean" was often used to describe trees and other graphical models, in fact close links were seen with registration, shape analysis, and other themes represented elsewhere in the program. The overall objective of the SAMSI AOOD program of bringing together hitherto disparate research areas was well realized both, within the Transition Workshop and in new research programs launched during the year.

1.4.2 Activities

The program for the workshop was as follows:

- **Functional Data Analysis** *Hans-Georg Müller, organizer; Jane-Ling Wang, chair*
  - Jeff Morris, Hierarchical methods for the analysis of object data
  - Hongxiao Zhu, Robust classification of functional and quantitative image data using functional mixed models
  - Sylvie Tchumtchou, Online variational Bayesian inference in hierarchical models for correlated high-dimensional data
  - Discussion: Jane-Ling Wang
- **Dynamics** *J.O. Ramsay, organizer and chair*
  - Jiguo Cao, Quantitative trait loci mapping with differential equation models
  - David Campbell, Parameter estimation from locally enforced differential equation models
  - Hulin Wu, High-dimensional ODEs for dynamic gene regulatory networks
  - Jim Ramsay, Reflections on impacts and issues for statistical methodology for dynamic models generated by the AOOD project
- **Trees** *J.S. Marron, organizer and chair*

- Steve Marron, Driving example, background and research overview
- Sean Skwerer, Phylogenetic trees and stickiness
- Dan Shen, Dyck path and branch length analysis
- Lingsong Zhang, Non-negative matrix factorization approach to tree analysis
- Yongdai Kim, Thread bridging example: Pseudo-Bayesian factor analysis
- John Aston, Thread bridging example: Phylogenetic trees with dialects as leaves D
- Discussion: J. S. Marron, John Aston, Lingsong Zhang and Hernando Ombao

- **Shapes and Manifolds** *I. Dryden, organizer and chair*
  - Victor Patrangenaru, Statistical analysis of object data
  - Sungkyu Jung, Principal nested shape spaces and an application to reduction of the number of landmarks
  - Sebastian Kurtek, Registration of functional data using the Fisher-Rao metric
  - Jingyong Su, Fitting optimal curves to time-indexed, noisy observations of stochastic processes on nonlinear manifolds
  - Ian Dryden, Metrics, manifolds and geometric correspondence
  - Discussion: Ian Dryden, J. S. Marron, Victor Patrangenaru

- **Brain Imaging** *J-L. Wang organizer, John Aston and Hernando Ombao, chairs*
  - Heipeng Shen, Hemodynamic response function modeling
  - Tingting Zhang, Nonparametric inference of hemodynamic response for fMRI data with inhomogeneous variances through kernel smoothing
  - Ci-Ren Jiang, Nonparametric response function estimation via FPCA with an application to dynamic PET data
  - John Aston, Spatial functional data, temporal sequences and populations of change points for fMRI analysis
  - Seonjoo Lee, Independent component analysis for autocorrelated sources with an application to fMRI
  - Discussion: John Aston, Ian Dryden, Jeff Morris, Hernando Ombao, Heipeng Shen, Jane-Ling Wang

2 Courses and Workshop for Students

Two one semester courses for graduate students and a workshop for undergraduate students were offered under the auspices of this program.

2.1 Graduate Courses

- *Analysis of Object Data I*

  Fall Semester 2010

  Principal Instructors: I. Dryden, J.S. Marron, H.G. Müller, J. O. Ramsay, J.L. Wang

Lectures were given at SAMSI on Wednesdays, 4:30-7:00 p.m.

This was the first of two courses associated with the SAMSI program on Analysis of Object Data, offered for graduate students at University of North Carolina, Duke University and North Carolina State University. These courses provided an introduction into selected areas of object oriented data analysis, aiming at the following topics:

1. Introduction to Object Oriented Data (Marron)
2. Shape Analysis and Related Topics (Dryden) Introduction to Statistical Shape Analysis, Non-Euclidean Shape Spaces, Distances and Shape Co-ordinates, Procrustes Analysis, Principal Components Analysis and Geodesics.
3. Functional Data (Mller) Introduction to Functional Data, Functional Regression Models, Time Warping, Empirical Dynamics.
4. Functional and Longitudinal Data (Wang) Functional Principal Component Analysis, Modeling with Covariates, Interface Functional and Longitudinal Data Analysis.
5. Functional and Dynamic Data (Ramsay)

- *Analysis of Object Data II*

  Spring Semester 2011

  Principal Instructors: J. S. Marron, J. O. Ramsay

  Lectures were given at SAMSI on Wednesdays, 4:30-7:00 p.m.

  This was the second graduate course associated with the SAMSI program on Analysis of Object Data. It covered topics different from those covered in the course taught in the Fall and consisted of three main segments:

  1. Dynamic systems: (Ramsay) Models for multivariate functional data that explicitly model change through the involvement of one or more derivatives in the model specification. Topics include the anatomy of a dynamic system, dynamic systems as extensions of functional models, and parameter estimation and inference for systems where analytical solutions are impossible.
  2. Manifold data: (Ramsay and Marron) This topic extends functional data analysis in a number of ways, where data are typically distributed over time and/or space, to situations where the data are distributed over manifolds embedded within higher dimensional spaces. Along with a quick review of classic subjects such as principal components analysis and test theory, more advanced topics include medial shape representations, diffeomorphisms in image analysis, and diffusion tensor imaging.
  3. Tree-structured data: (Marron) This section contrasts the very diverse combinatorial, folded Euclidean and Harris correspondence approaches for this new area of data analysis.

2.2 Undergraduate Workshop

A two-Day Undergraduate Workshop was held at SAMSI on February 25-26, 2011, with a teaching program organized by Hans-Georg Müller. This workshop was part of SAMSI's Education and Outreach Program for 2010-2011.

The focus was the topic *Analysis of Object Data*, and the following lectures and events took place:

- Pierre Gremaud, N.C. State University and SAMSI, Welcome and Introduction
- Jim Ramsay, McGill University, Introduction to Analysis of Object Data
- Ci-Ren Jiang, SAMSI, MATLAB demo
- Snehalata Huzurbazar, University of Wyoming, Exmples of Collaborative Research Projects which use Analysis of Object-Oriented Data
- Hulin Wu, University of Rochester Dynamics, Modeling as a Weapon to Defend Ourselves Against Threats from Infectious Diseases and Bioterrorist Attacks
- Junheng Ma, SAMSI, R introduction
- Hongxiao Zhu, SAMSI , Introduction to Statistical Analysis of Functional Data
- Snehalata Huzurbazar, University of Wyoming, and Sylvie Tchumtchoua, SAMSI, MATLAB lab
- Pierre Gremaud, N.C. State University and SAMSI, Career Options
- Yolanda Muñoz-Maldonado, Michigan Tech, Sample Size Calculation for Functional Data
- David Degras, SAMSI, Topics in Functional Data Analysis
- Yolanda Muñoz-Maldonado, Michigan Tech and David Degras, SAMSI R lab

3 Report for the Working Groups on Functional Data Analysis

Leaders: Jeffrey Morris (University of Texas, Houston), Hans-Georg Müller (University of California, Davis), Jane-Ling Wang (University of California, Davis), Fang Yao (University of Toronto) and Hao Helen Zhang (North Carolina State University)

Functional data objects have been studied for quite some time now. In recent years, interest in this area has substantially increased, with a constant influx of new researchers and also established researchers starting to work in this area. Among the various types of object data, functional data are the most Euclidean and the main complication in their analysis stems from their infinite-dimensional nature, so that these objects lie in an infinite-dimensional Hilbert space. Moreover, in practice, the sample of random functions is usually not fully observed. Rather, observations are often sparse and noisy, especially when the functional data are observed in longitudinal studies. There are often complex dependence structures that one needs to deal with. Problems such as functional clustering and classification are not fully understood at this time. In addition, theoretical problems arise in function spaces, such as the non-existence of a Lebesgue density and these require creative solutions.

The mix of applied and theoretical challenges one faces in Functional Data Analysis at the cross-roads of stochastic processes, functional analysis, multivariate analysis and longitudinal and hierarchical modeling requires teams of researchers that bring mathematical, computational and various areas of statistical expertise to the table, and the SAMSI environment proved excellent for addressing some of the key issues. Two major working groups formed in this area, one focusing on more general topics and especially regression and classification problems for functional data (Leaders: Fang Yao and Helen Hao Zhang), and the other one on hierarchical and Bayesian functional modeling (Leader: Jeff Morris). Both groups shared a sizeable number of members and gained from mutual interactions and discussions between members. In addition, a workshop was held that emphasized the interface between longitudinal and functional data analysis. Details about this workshop can be found in the Workshops section of the report

The research in Functional Data Analysis profited from the Shape and Manifolds thread in individual discussions and joint membership in working groups. There were also close ties with the Dynamics thread and especially with the Brain Imaging thread, as functional data analysis is immediately applicable for many problems in Brain Imaging. In turn, many challenging problems for research on functional data analysis are motivated by the need to analyze various types of brain images, and both of these threads profited immensely from exchanges and cross-interactions between their respective members.

3.1 Report for the Working Group: Statistical Inference for Functional Data

Leaders: Fang Yao (University of Toronto) and Hao Helen Zhang (North Carolina State University)

3.1.1 Topics

Functional data pose unique challenges to statistical inferences because of their high-dimensional and complex structure. A further complication are sparse and noisy observations of the underlying random functions. The working group focused on two main goals: (1) to study inferential properties of existing approaches for functional data analysis (FDA) and to understand their fundamental behaviors; (2) to develop new, flexible, and powerful tools for FDA, including regression, classification, clustering, principal component analysis, and dimension reduction for such data. Situations of both densely and sparsely sampled functional data were considered. A variety of particularly relevant research topics and important future issues were identified and became topics for research conducted by group members. These topics include variable selection, model selection, sparse estimation, robust estimation, and experiment design for functional data. The new methods are motivated from real-world problems and have important applications in various fields such as image data, speech data, and shape analysis, for which connections with other workgroups proved useful. Concepts that were studied by group members also included time warping and manifolds with connections to other workgroups. Furthermore, extensions to user-friendly software and packages were discussed and partly implemented.

3.1.2 Participants

Participants in the work group were mostly new and junior researchers. The participants included SAMSI visitors, postdoctoral fellows, graduate students, local faculty and scientists: John Aston, Matt Avery, Graciela Boente, Herve Cardot, Jeng-Min Chiou, David Degras, Jimin Ding, Pang Du, Xingdong Feng, Kaushik Ghosh, Jinjiang He, Giles Hooker, Sungkyu Jung, Seonjoo Lee, Lexin Li, Yufeng Liu, Wenbin Lu, Junheng Ma, Hans-Georg Müller, Yolanda Munoz, Todd Ogden, Juhyun Park, Philip Reiss, Richard Samworth, Damla Şentürk, Haipeng Shen, Joon Jin Song, Ulrich Stadtmüller, Wenwen Tao, Sylvie Tchumtchoua, Haonan Wang, Jane-Ling Wang, Judy Wang, Liwei Wang, Yishi Wang, Hulin Wu, Shuang Wu, Yichao Wu, Fang Yao, Nuen Tsang Yang, Bo Zhang, Hao Helen Zhang, Lingsong Zhang, Jun Zhang, Hongxiao Zhu, Frank Zou, Jian Zou.

3.1.3 Activities and Research

The working group met throughout the entire AOOD program during 2010-2011 with weekly group discussions and presentations. Several subgroups were formed, which focused on specific topics. Close collaborations were established among the group participants. Group members were actively engaged in a variety of activities, including teaching special topic courses, group meetings, talks and presentations, and general as well as research discussions. Major activities and achievements in chronological order were as follows:

In September 2010, the working group was formed at the AOOD Opening workshop. Several important topics were identified in the area of statistical inference:

- Variable selection for functional linear regression models (FLM)
- Model selection for functional additive models (FAM)
- Classification and clustering for functional data
- Extensions of functional principal component analysis
- Experimental design for functional data
- Domain selection for functional data

From September to November of 2011, the entire working group met weekly to discuss important issues, study areas of general interest and to address theoretical and computational challenges. Each meeting focused on one topic and started with a couple of talks about the background for the problem, aiming to put all the participants on the same page. Existing works and open problems were then discussed, which stimulated the participant's interest to work on these problems and to establish collaborations with each other.

During November 8-10, 2010, a workshop on the Interface between Longitudinal and Functional Data was held at SAMSI under auspices of this working group, organized by Marie Davidian, Hans-Georg Müller and Fang Yao (for more details, see Section 1.2).

In Spring 2011, several small research groups were formed through interactions between participants, each focused on a narrower topic with the goal to do research on the topic. These small focused groups met regularly to collaborate on the selected problem, share research findings and progress, and discuss relevant issues. Manuscripts, software, and research papers were prepared or submitted. Some more detailed descriptions of the activities of a few such focus groups follow. These groups are a selection from a much larger number of such groups that were formed in Fall 2010; see the more comprehensive list of papers below, where the results of nearly all such groups are collected.

1. Variable Selection in Semiparametric Functional Linear Models

   Participants: Dehan Kong, Fang Yao and Hao Helen Zhang

   The goal of this group was to study the variable selection problem for semiparametric functional linear models. The group developed a new class of semiparametric functional regression models for jointly modeling the functional and non-functional predictors, identifying important scalar covariates while taking into account functional covariates. The new method takes advantage of modern shrinkage technique to achieve sparse estimation for parametric terms. The algorithm for this method is convenient for implementation and very efficient in identifying important scalar covariates. A preprint of the paper has been prepared.

2. Model Selection for Functional Additive Models

   Participants: Hongxiao Zhu, Fang Yao, Hao Helen Zhang

   Nonparametric functional data regression provides a flexible alternative to linear models in functional data analysis. This group focused on model selection in this setting, which is much more challenging than functional linear models, due to the unspecified form of effects for the functional predictors. The group proposed a new type of regularization framework which can automatically select important predictors and hence build sparse estimation models. Large sample properties of the new estimator are studied.

3. Experimental Design & Survey Sampling for Functional Data

   Participants: David Degras, Yolanda Munoz Maldonado, Herve Cardot

   The project deals with survey sampling techniques for functional data. The goal is to develop time-varying sampling schemes that improve statistical inference in finite populations. This type of work has important applications in sensor networks where huge numbers of signals can be observed at fine time scales but full observation and/or analysis of the data are operationally impossible. A good example is to estimate the electricity consumption in a large population based on digital meter readings. The group proposed two time-varying stratified sampling schemes: partial or full resampling. Population mean function was estimated by the Horvitz-Thompson (HT) survey estimator. The research activities included: (1) to determine the bias and covariance of the HT estimator, and comparisons with time-invariant samples; (2) to establish asymptotic theory and determine conditions on sampling rate, resampling frequency, and renewal rate needed for convergence results; (3) numerical studies with real and simulated data sets.

4. Functional Data Classification and Clustering

   Participants: Yichao Wu, Yufeng Liu, Junheng Ma, Fang Yao, Hao Helen Zhang

   The group has been working on two projects, respectively focusing on functional data classification and clustering. In the first project, for data with a functional predictor and a categorical response, the group proposes functional robust support vector machines. With the aid of functional principal component analysis provided by the PACE package, functional robust support vector machines can accommodate sparse and irregular functional data. One paper has been submitted on this project. In the second project, a new weighted distance is proposed for functional data clustering based on principal component analysis.

5. Domain Selection for Functional Data

   Participants: Giles Hooker and Damla Senturk

   The problem of domain selection belongs to an important class of problems and motivated from real-world problems. Typical examples where domain selection is crucial include: predicting the lifespan of medflies from their fecundity, estimating the length of hemodynamic response function, and vehicle exhaust measurement dependence of an engine function. The group proposed a profiled estimate and studied its theoretical properties including consistency and CLT. For practical implementation, various tuning methods were studied and compared.

6. Robust Functional Principal Components

   Participants: Graciela Boente and Jane-Ling Wang

   In many situations, data are recorded over a period of time and may be regarded as realizations of a stochastic process. In this paper, robust estimators for the principal components are considered by adapting the projection pursuit approach to the functional data setting. Our approach combines robust projection–pursuit with different smoothing methods. Consistency of the estimators are shown under mild assumptions. The performance of the classical and robust procedures are compared in a simulation study under different contamination schemes.

   This is a long term project that began many years ago but due to the different continents the collaborators reside, the progress has been slow. Both Boente and Wang visited SAMSI in

Fall 2010 and completed the project. Boente gave a presentation of the results in a SAMSI workshop "Functional approaches for longitudinal data" in November, 2010 and a paper was submitted shortly afterwards to the Annals of Statistics. They have received an invitation for a revision. The revision is still under review at this time.

7. Varying coefficient Cox model with functional and longitudinal covariates

   Participants: Jiming Ding and Jane-Ling Wang

   Ding and Wang worked on this project while they were both visiting SAMSI in Fall 2010. The project involves joint modeling of survival data and functional/longitudinal covariates. A varying coefficient Cox model is assumed for the survival data and nonparametric mixed-effects model is used to model the functional/longitudinal covariates in the Varying coefficient Cox model. All unknown components/quantities are modeled nonparametrically through a joint nonparametric likelihood and their asymptotic properties investigated. We have completed all theory and numerical analysis and have started to write a draft paper to be submitted to the Annals of Statistics. The plan is to have a preprint in a few months.

3.1.4 Additional Activities

- Participation in the Opening Workshop, Sep 12-14, 2010
  - Hans Müller and Jane-Ling Wang were co-organizers
  - Fang Yao gave Tutorial Lecture 1: Functional Data Analysis and Related Topics. (9:00-10:30am, Sep 12, 2010)
- Participation in the SAMSI-AOOD course:
  - Hans Müller gave several lectures for "Analysis of Object Data I", Fall 2010
  - Jane-Ling Wang gave several lectures for "Analysis of Object Data I", Fall 2010
- Participation in the AOD Workshop: Interface Functional and Longitudinal Data Analysis - November 8-10, 2010 (for more details see separate report below)
  - Hans Müller and Fang Yao served as co-organizers
  - Damla Senturk gave an invited talk "Functional Varying Coefficient Models for Longitudinal Data" (10:00-10:20am, Nov 8, 2010)
  - Fang Yao gave an invited talk "Additive Modeling of Functional Regression and its Gradients" (3:30-4:00pm, Nov 8, 2010)
  - Hao Helen Zhang gave an invited talk "Variance Component Selection in Linear Mixed Models" (10:00-10:30am, Nov 9, 2010)
  - Graciela Boente gave an invited talk "Robust Functional Principal Components: a projection-pursuit approach" (11:15-11:45am, Nov 9, 2010)
  - Jimin Ding gave an invited talk "Time-varying Coefficient Cox model with Nonparametric Longitudinal Covariates" (9:05-9:25am, Nov 10, 2010)
- Participation in the AOOD transition workshop, June 9-11, 2011
  - Hans Müller and Jane-Ling Wang served as co-organizers
  - Hongxiao Zhu gave an invited talk "Robust Classification of Functional and Quantitative Image Data using Functional Mixed Models" (9:30-9:50am, June 9, 2011)

– Jane-Ling Wang gave an invited talk "Discussion and Connections to FDA Working Groups" (10:10-10:40am, June 9, 2011)
– Yichao Wu gave an invited talk "Continuously Additive Models for Functional Regression Analysis" (11:10-11:40am, June 9, 2011)
– David Degras gave an invited talk "Longitudinal Survey Methods for Functional Data" (11:40-12:00pm, June 9, 2011)
– Jane-Ling Wang served on the panel for "Discussion and Future Developments" (12:00-12:30pm, June 9, 2011)

3.1.5 Working Papers and Publications

The works listed in the following benefitted from our work group at SAMSI in one or several of the following ways: (1) research inspired by work group discussions or workshops; (2) research resulting from collaborations among work group participants that were begun during participant's stay at SAMSI; (3) research continued or completed (submitting or revising papers) during participant's stay at SAMSI.

Published/Accepted

1. Acar, E., Craiu, R.V., Yao, F. (2011). Dependence calibration in conditional copula: a nonparametric approach. *Biometrics*, accepted.

2. Chen, K., Chen, K., Müller, H.G., Wang, J.L. (2011). Stringing high-dimensional data for functional analysis. *Journal of American Statistical Association* **106**, 275–284.

3. Chen, D., Hall, P., Müller, H.G. (2011). Single and multiple index functional regression models with nonparametric link. *Annals of Statistics* **39**, 1720–1747.

4. Wu, S., Müller, H.G. (2011). Response-adaptive regression for longitudinal data. *Biometrics*, accepted

5. Chen, K., Müller, H.G. (2011). Conditional quantile analysis when covariates are functions, with application to growth data. *J. Royal Statistical Society B*, accepted

6. Ding, J.-M., Symanzik, J., Sharif, Wang, J. -L., Duntley, Shannon, W. (2011). Powerful Actigraphy Data Through Functional Representation. *Chance*, accepted.

7. Park, S. Y. and Liu, Y. (2011). Robust penalized logistic regression with truncated loss. *The Canadian Journal of Statistics*, accepted.

8. Li, P.L. and Chiou, J.M. (2011) Identifying cluster numbers for subspace projected functional data clustering. *Computational Statistics and Data Analysis* **55**, 2090-2103.

9. Liu, Y. and Yuan, M. (2011). Reinforced multicategory support vector machines. *Journal of Computational and Graphical Statistics*, accepted.

10. Liu, Y., Zhang, H. H., and Wu, Y. (2011). Soft or hard classification? Large margin unified machines. *Journal of the American Statistical Association* **106**, 166-177.

11. Lu, W., Zhang, H. H., and Zeng, D. (2011). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, accepted.

12. Müller, H.G. (2011). Functional data analysis. *International Encyclopedia of Statistical Science*, Ed. Lovric, M. Springer Science Business Media, Heidelberg. (Extended version available in StatProb: The Encyclopedia Sponsored by Statistics and Probability Societies, id 242).

13. Müller, H.G., Sen, R., Stadtmüller, U. (2011). Functional data analysis for volatility. *Journal of Econometrics*, acceped

14. Wu, Y. (2011). An ordinary differential equation-based solution path algorithm . *Journal of Nonparametric Statistics*, **23**, 185-199.

15. Wu, Y. and Li, Lexin (2011). Asymptotic Properties of Sufficient Dimension Reduction with A Diverging Number of Predictors. *Statistica Sinica* **21**, 707-730.

16. Wu, Y. and Liu, Y. (2011). Non-crossing large-margin probability estimation and its application to robust SVM via preconditioning. *Statistical Methodology* **8**, 56-67.

17. Yang, W., Müller, H.G., Stadtmüller, U. (2011). Functional singular component analysis. *J. Royal Statistical Society B* **73**, 303–324.

18. Yao, F., Fu, Y., Lee, T.C.M. (2011). Functional mixture regression. *Biostatistics*, **12**, 341-353.

19. Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, accepted.

20. Zhang, Z., Müller, H.G. (2011). Functional density synchronization. *Computational Statistics and Data Analysis* **55**, 2234–2249.

21. Zhu, H., Brown, P. J., Morris, J. S. (2011). Robust, adaptive regression in functional mixed models framework. *Journal of the American Statistical Association*, accepted.

Submitted/In Revision

1. Ahn, M., Zhang, H. H., and Lu, W. (2011) Moment-based method for random effect for selection in linear mixed models. *Statistica Sinica*, revised.

2. Bosca, L., Boente, G., Tyler, D. and Wang, J.-L. (2011) . Robust functional principal components. *Annals of Statistics*, revised.

3. Cardot, H., Degras, D., and Josserand, E. (2011). Confidence bands for Horvitz-Thompson estimators using sampled functional data. Submitted.

4. Chen, D., Müller, H.G. (2011). Nonlinear manifold representations for functional data. Submitted.

5. Ding, A.A. and Wu, H. (2011). Estimation of ODE parameters using constrained local polynomial regression. Submitted.

6. Fang, Y., Wu, H., Zhu, L. (2011). A Note on Data Augmentation-Based Pseudo-Least Squares Estimation for ODE Models. Submitted.

7. Müller, H.G., Wu, Y. Yao, F. (2011). Continuously additive models for functional regression. Submitted.

8. Şentürk, D., Ghosh, S. and Nguyen, D. V. (2011). Exploratory time varying lagged regression for longitudinal data. *Biometrics*. Submitted.

9. Wei, F. and Zhu, H. (2011). Group Coordinate Descent Algorithms for Nonconvex Penalized Regression. *Computational Statistics & Data Analysis*. Submitted.

10. Wu, Y. and Liu, Y. (2011). Functional robust support vector machines for sparse and irregular longitudinal data. Revised.

11. Wu, H., Miao, H., Xue, H., Topham, D.J., and Zand, M. (2011). Quantifying immune response to influenza virus infection via multivariate nonlinear ODE models with partially observed state variables and time-varying parameters. *Journal of American Statistical Association*, submitted.

12. Wu, H., Xue, H., Kumar A. (2011), Numerical algorithm-based estimation methods for ODE models via penalized spline smoothing, *Biometrics*, revised.

13. Zhu, H., Brown, P. J., Morris, J. S. (2011). Robust Classification of Functional and Quantitative Image Data using Functional Mixed Models. *Biometrics*, revised.

In Preparation/Preprint

1. Asencio, M., Hooker, G. and Gao, H. (2011). Functional Convolution Models. *Preprint.*

2. Avery, M, Zhang, H. H. , Wu, Y. (2011) . Sparse estimation in functional data analysis. *In preparation.*

3. Chen, Y. and Samworth, R. J. (2011). Smoothed log-concave maximum likelihood estimation with applications. *Preprint.*

4. Chiou, J.M., Ma, Y. and Tsai, C.L. (2011). Functional time-varying random effects models for longitudinal data. *Preprint.*

5. Chiou, J.M. and Müller, H.G. (2011). Linear manifold modeling of multivariate functional data, with application to traffic flow analysis. *Preprint.*

6. Degras, D. (2011). Longitudinal survey sampling for functional data. *In preparation.*

7. Ding, J.-M. and Wang, J.-L. (2011). Varying coefficient Cox model with functional and longitudinal covariates. *In preparation.*

8. Hooker, G.. (2011) Domain selection for functional linear models. *In preparation.*

9. Kong, D. , Yao, F., Zhang, H. H. (2011). Semiparametric functional linear model. *In preparation.*

10. Lee, W. and Liu, Y. (2011). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Preprint* .

11. Li, L., Yao, F., Craiu, R. V. (2011). Minimum description length principle for correlated data. *Preprint.*

12. Li, L., Zhou, H., and Zhu, H. (2011). Tensor regression with applications in neuro-imaging data analysis. *In preparation.*

13. Li, N., and Zhang, H. H. (2011) Sparse learning in multi-class classification. *Preprint.*

14. Lin, C. Y., Zhang, H. H., Bondell, H., and Zou, H. (2011) Nonparametric variable selection in quantile regression. *Preprint.*

15. Liu, C., Ray, S., Hooker, G. and Friedl, M. (2011) . Functional factor analysis for periodic remote sensing data. *Preprint.*

16. McLean, M., Hooker, G., Saicu, A. and Ruppert, D. (2011) Functional generalized additive models. *Preprint.*

17. Du, P., Cheng, G., and Zhang. H. H. (2011) . Structure selection for nonparametric survival analysis. *In preparation.*

18. Samworth, R. J. (2011). Optimal weighted nearest neighbor classifiers. *Preprint.*

19. Şentürk, D. (2011). Efficient estimation for generalized varying coefficient models with longitudinal data. *In preparation.*

20. Shah, R. and Samworth, R. J. (2011), Variable selection with error control: Another look at stability selection. *Preprint.*

21. Staicu, A. M., Şentürk, D., Carroll, R. J. (2011). Generalized time-varying spatial regression of multilevel functional data. *In preparation.*

22. Wang, L., Wu, Y., and Li, R. (2011). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Preprint.*

23. Wong, R.K.W., Yao, F., Lee, T.C.M. (2011). Robust estimation for generalized additive models. *Preprint.*

24. Wu, Y. (2011). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Preprint* .

25. Zhou, H. and Wu, Y. (2011) A general path algorithm for regularized statistical estimation. *Preprint.*

26. Zhu, H., Dunson, D. B. (2011). Bayesian graphical models for multivariate functional data. *Preprint.*

27. Zhu, H., Yao, F., Zhang, H. H. (2011). Component Selection in Functional Additive Models. *Preprint.*

3.2 Report for the Working Group: Hierarchical Modeling

Leader: Jeffrey S. Morris (University of Texas, Houston)

3.2.1 Goals

This working group is interested in developing hierarchical modeling approaches for object data, including functions, images, and more general structures like shapes and trees. The goal is to develop inferential methodology motivated by specific applications yielding complex, structured data. The idea of hierarchical modeling implies flexible, unified models that can simultaneously take into account variability and structure from multiple sources in the data set, within and between objects, and induced by the design or other measured covariates. Both Bayesian and frequentist approaches have been considered, and discussion of the connections and distinctions among existing Bayesian and frequentist approaches in the literature done. As longer term goals, our objective is to build new unified methodology for modeling and performing inference for functional, image, shape, and other object data using as building blocks various modeling tools that have been used in modeling in simpler contexts, but not yet used together. We will also make data and software available for these methods so the work can have impact both in terms of stimulating more methodological research and producing tools investigators can use to analyze their object data.

3.2.2 Participants

Jeffrey S. Morris, Syvie Tchumtchaoua, Jaeyong Lee, Jim Berger, Susan Bayarri, Ci-Ren Jiang, Xia Wang, Brian Reich, Howard Bondell, Ana-Maria Staicu, Veera Baladandayuthapani, Michele Guindani, Todd Ogden, Philip Reiss, Scott Holan, Josue Martinez, Hongtu Zhu, Raymond Carroll, Giles Hooker, Ian Dryden, Darren Wilkenson, Dubois Bowman, Hongxiao Zhu, Xiaojing Wang, Sarat Dass, Genevera Allen, Xiaojing Wang, David Dunson, Jiguo Cao.

3.2.3 Activities

Throughout the year, we had a number of presentations and discussions involving various types of hierarchical modeling approaches that have been applied and/or can potentially be applied to the analysis of object data, including the following:

- Bayesian Variable Selection Methods
- Sparsity Priors and Penalties
- Multilevel Functional Principal Components
- Functional Mixed Models
- Bayesian Nonparametrics
- Density and Quantile Regression
- Spatial Models for Dependency
- Independent Components Analysis
- Generalized Least Squares Matrix Decomposition
- Product Kernels

These talks were done weekly throughout the fall, and then there were a few more in the spring. These talks provided general knowledge to all attendees, and sparked interesting conversations as well as new collaborations, which spun off the group and worked on research manuscripts. Some of these new collaborations have already led to manuscripts that have been submitted, and others that are in preparation (see list below).

In the spring, we split the working group into three subgroup areas and continued to meet every other week as subgroups. The groups and their defined aims were as follows:

1. FDA and Bayesian Nonparametrics: Develop methods applying various Bayesian nonparametric modeling approaches to the analysis of functional and other object data.

2. Object Regression: Develop flexible methods and models for regression analysis for object data such as functional, image and shape data.

3. Hierarchical Methods for Image Analysis: Develop hierarchical methods for the analysis of large-scale image data sets, including biomedical imaging data and spatial data.

During our subgroup meetings, we had some pre-arranged presentations, but also brainstormed research ideas and reviewed literature in various areas with the goal of stimulating new collaborations and research. Following are some of the projects within these areas, which include some projects that have been completed, some with substantial progress, and some that have just been started.

- FDA and Bayesian Nonparametrics
  - Development of Bayesian nonparametric approaches to analyze functional and image data.
  - Development of a general theoretical framework for Bayesian nonparametric functional data analysis, describing important support properties in this framework, assessing which existing methods possess these properties, and use these as motivations for developing new methods.
  - Develop methods to perform Functional density regression, which model covariate effects on the entire distribution of the function, including mean, covariance, and extreme quantiles.
  - Detect change points using functional Dirichlet Processes.
- Object Regression
  - Develop methods incorporating local variable selection in nonparametric regression and functional data settings, using priors with positive probability given to flat regions of the curves.
  - Develop method to analyze longitudinal shape data using functional mixed models.
  - Develop methods for fitting functional mixed models using adaptively regularized sparse functional principal components, and using hybrid transforms that include both global and local components.
  - Develop methods to model functional data with spatially heterogeneous shape characteristics.
  - Develop sufficient dimension reduction methods, which aim to identify the smallest set of linear combinations of the predictors that retain all information in the predictors about the response distribution, and apply the new methods to develop climate indices for forecasting yearly hurricane intensity.
  - Investigate the performance of functional principal components in settings with complex, high dimensional functional and image data, to provide guidance on when these models should be used and how many dimensions to model.

  - Develop hierarchical methods to model spatially correlated functional and image data.
  - Develop new method for regression of scalar responses on high dimensional functional binary predictors, and apply to the analysis of GWAS data.
  - Write review chapter on multi-level functional data analysis methods.
  - Develop method for analysis of nonstationary time series using functional mixed models involving spectrograms and wavelet transforms.
  - Develop method to perform robust functional inference for spatially heterogeneous functional and image data.
  - Develop methods to classify individuals based on functional data observed in nested hierarchical structure.
  - Develop methods to analyze functional data observed on a spatial grid.
  - Develop methods for analyzing spatially correlated binary longitudinal data within a multilevel model.

- Hierarchical Methods for Image Analysis
  - Develop hierarchical model-based methods to simultaneously perform subject- and group-level analyses of complex, extremely high dimensional quantitative image data (e.g. fMRI) using functional mixed models.
  - Develop new methods for multivariate longitudinal data using flexible semiparametric Bayes models.
  - Motivated by massive dimensional dynamic functional data, develop flexible online Bayesian methods, which work by reading in a slice of data at one time and approximating the posterior based on these data, and then updating the approximation as additional data are read in.
  - Develop methods to identify genetic pathways that affect a subject's functional connectivity in response to a series of external stimuli, using a hierarchical Bayesian model combining information across fMRI and genetic data.
  - Develop Bayesian Independent Components Analysis (ICA) methods that exploit prior information about spatial association and known connectivity patterns between anatomic features.

3.2.4 Additional Activities

- Participated in longitudinal/functional data workshop, with the following invited talk:
  - 2:30-3:00 Jeffrey S. Morris, University of Texas MD Anderson Cancer Center Adaptive, Robust Functional and Image Regression in Functional Mixed Models
- Participated in transitional workshop, with the following talks presented:
  - 9:00-9:30 Jeffrey S. Morris, MD Anderson Hierarchical Methods for Analysis of Object Data
  - 9:30-9:50 Hongxiao Zhu, SAMSI Robust Classification of Functional and Quantitative Image Data using Functional Mixed Models
  - 9:50-10:10 Sylvie Tchumtchou, SAMSI Online Variational Bayesian Inference in Hierarchical Models for Correlated High-dimensional Data

  - 12:00-12:30 Discussion and Future Developments: Steve Marron, UNC; Jeff Morris, MD Anderson; Jim Ramsay, McGill University; Jane-Ling Wang, University of California, Davis

- Organized a Topic Contributed Session for 2011 Joint Statistical Meetings in Miami Beach, Florida, as follows: Hierarchical Methods for Object Data: SAMSI Object Data Program Working Group Organizer(s): Jeffrey S. Morris, The University of Texas MD Anderson Cancer Center Chair(s): Denis Larocque, HEC Montréal
  - 10:35 AM Bayesian Hierarchical Functional Models for High-Dimensional Genomics Data Veera Baladandayuthapani, The University of Texas MD Anderson Cancer Center
  - 10:55 AM Classification of Unknown Powders Using a Support Vector Machine Classification Model - Jessi Cisewski, The University of North Carolina at Chapel Hill ; Jan Hannig, The University of North Carolina at Chapel Hill ; Emily Snyder, Environmental Protection Agency
  - 11:15 AM Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Datatypes - Eric Frazer Lock, The University of North Carolina at Chapel Hill
  - 11:35 AM Estimating Shape-Constrained Functions Using Gaussian Processes - Xiaojing Wang, Duke University ; James Berger, Duke University
  - 11:55 AM An Application of Nonparametric Function Estimation for Planning Reconstructive Surgical Procedures for the Skull - Daniel Osborne, Florida State University ; Victor Patrangenaru, Florida State University ; Xiuwen Liu, Florida State University ; Hillary W. Thompson, Louisiana State University
  - 12:15 PM Floor Discussion

3.2.5 Publications

* Also associated with FDA Inference working group ** Also associated with Brain Imaging working group

Published/Accepted:

1. Cao, J., Cai, J. and Wang, L. (2011). Estimating Curves and Derivatives with Parametric Penalized Spline Smoothing, Accepted by *Statistics and Computing.*

2. Dass, S. C., Lim, C. Y. and Maiti, T. (2011). Default Bayesian Analysis for Multivariate Generalized CAR Models, To appear in *Statistica Sinica.*

3. Lim, C. Y. and Dass, S. C. (2011). Assessing Fingerprint Individuality Using EPIC: A Case Study In The Analysis Of Spatially Dependent Marked Processes,*Technometrics*, vol. 53, no. 2, pp. 112-124.

4. **Morris JS, Baladandauthapani V, Herrick RC, Sanna PP, and Gutstein HG (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomic data, *Annals of Applied Statistics*, to appear.

5. *Staicu, A. M., Crainiceanu, C., Reich, D. and Ruppert, D. (2011). Modeling functional data with spatially heterogeneous shape characteristics, *Biometrics*, to appear.

6. Wu, R., Cao, J., Huang, Z., Wang, Z., Gai, J. and Vallejos, E. (2011). Systems mapping: how to improve the genetic mapping of complex traits through design principles of biological systems, *BMC Systems Biology* 5:84, 1-24.

7. Zhu, H., Brown, P. J., and Morris, J. S. (2011). Robust, Adaptive Functional Regression in Functional Mixed Model Framework. *JASA*, to appear.

Submitted/In Revision:

1. Crainiceanu, C. M., Caffo, B. S., and Morris, J. S. (2011). Multilevel Functional Data Analysis. Submitted to *SAGE Handbook for Multilevel Modeling.*
2. Dass, S. C., Lim, C. Y and Maiti, T. (2011). Detecting Change Points of Cancer Incidence Rates using Functional Dirichlet Processes. *Submitted.*
3. Li, J.J., Jiang, C.R., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Sparse linear modeling of RNA-seq data for isoform discovery and abundance estimation. In revision for *PNAS.*
4. Martinez, J. G., Bohn, K. M., Carroll, R. J. and Morris, J. S. (2011). A study of Mexican free-tailed bat syllables: Multi-domain modeling of nonstationary time series with high frequency content using Bayesian functional mixed models. *Under revision.*
5. *McLean, M. W., Hooker, G., Staicu, A. M., and Ruppert, D. (2011). Functional Generalized Additive Models. *Submitted.*
6. Montagna, S., Tokdar, S.T., Neelon, B. and Dunson, D.B. (2011). Bayesian latent factor regression for functional and longitudinal data. *Submitted.*
7. Morris, J. S. (2011). Statistical Methods for Proteomic Biomarker Discovery using Feature Extraction or Functional Data Analysis Approaches. Under Revision for *Statistics and its Interface.*
8. Wang, L. and Cao, J. (2011). Estimating Delay Differential Equations. Under revision for *Journal of Agricultural, Biological, and Environmental Statistics.*
9. *Staicu, A. M., Lahiri, S., and Carroll, R. (2011). Tests of significance for spatially correlated multilevel functional data. Submitted.
10. Tchumtchoua, S. (2011). "Bayesian Semiparametric Functional Generalized Linear Models for Longitudinal Data". *Submitted.*
11. Zhang, L., Baladandayuthapani,V., Mallick, B. K., Thompson, P., Bondy, M. and Do, K.A. (2011). Bayesian hierarchical structured variable selection methods with application to MIP studies in breast cancer. *Under revision.*
12. Zhu, H., Brown, P. J., and Morris, J. S. (2011). Robust classification of functional and quantitative image data using functional mixed models. *Submitted.*

In Preparation:

1. Baladandayuthapani, V., Morris, J.S., Coombes, K. R. and Abruzzo, L. (2011). Bayesian Adaptive Functional Linear Models for Copy Number Data. *In preparation.*
2. **Baladandayuthapani, V., Bharath, K., Baggerly, K., Czerniak, B., and Morris, J. S. (2011). Bayesian Spatial-functional Models for High-throughput genomic data. *In preparation.*

3. Cai, J. and Cao, J. (2011). Estimating Damage Accumulation Dynamic Models Hierarchical Bayesian Methods for Parameter Estimation in Dynamic Duration of Load Models. *In preparation.*

4. Chkrebtii, O. and Cao, J. (2011). Generalized Additive Modeling of Spatio-Temporal Trends in Salmon Productivity. *In preparation.*

5. Jiang, C.R., Aston, J.A. and Wang, J.L. (2011). Nonparametric response function estimation via FPCA with an application to dynamic PET data. *In preparation.*

6. Jiang, C.R. and Morris, J.S. (2011). Choosing number of principal components in high dimension, low sample size situations for complex functional data. *In preparation.*

7. Morris, J. S. and Allen, G. (2011). Analysis of fMRI data using functional mixed models and sparse empirically determined basis function representations. *In preparation.*

8. Pati, D. and Dunson, D.B. (2011). Bayesian closed surface fitting through tensor products. *In preparation.*

9. Serban, N., Staicu, A. M., and Carroll, R. J. (2011). Multilevel Spatially Correlated Binary Longitudinal Data. *In preparation.*

10. Staicu, A. M., Senturk, D., and Carroll, R. J. (2011). Generalized Time-Varying Spatial Regression of Multilevel Functional Data. *In preparation.*

11. **Tchumtchoua, S., Dunson, D. B., and Morris, J. (2011). Online Variational Bayes Inference for High-Dimensional Correlated Data.*In preparation.*

12. Tchumtchoua, S., Dunson, D. B., and Morris, J. (2011). A Heterogeneous Dynamic Structural Equation Model With Application to Brain Connectivity. *In preparation.*

13. Wang, Xia, Chen, M.-H., Dey, D. K., and Kuo C.-Y. (2011). Space-Time Modeling of Atlantic Cod Abundance in the Gulf of Maine. *In preparation.*

14. Wang, Xia and Sedransk, N. (2011). Bayesian Models on Biomarker Discovery Using Spectral Count Data in the Label-Free Shotgun Proteomics. *In preparation.*

15. Wang, Xia, Sedransk, N. and Tabb, D. (2011). Variability of Base Peak Intensities in Shotgun Proteomics Experiment: Perspectives from Functional Data Analysis. In preparation.

16. Wang, Xiaojing and Berger J. (2011). Estimated shape constrained functions using Gaussian Processes. *In preparation.*

17. Wei, F. and Zhu, H. (2011). Group coordinate descent algorithms for nonconvex penalized regression. *In preparation.*

18. Zhu, H. and Dunson, D. (2011). Bayesian graphical models for multivariate functional data. *In preparation.*

19. Zhu, H. and Morris, J. S. (2011). Functional Mixed Models for Serially Correlated Functional Data. *In preparation.*

4 Report for the Working Group: Dynamics and Inference

Leaders: David Degras (SAMSI), Giles Hooker (Cornell University) and James Ramsay (McGill University)

4.1 Participants

Karthik Bharath, Nicolas Brunel, David Campbell, Jiguo Cao, Oksana Chkrebtii, Sy-Miin Chow, Sarat Dass, Jimin Ding, Paul Fearnhead, Xingdong Feng, Kaushik Ghosh, Mark Girolami, Andrea Gottlieb, Serge Guillas, Snehalata Hurzurbazar, Hachem Kadri, Hamid Krim, Jaeyong Lee, Seonjoo Lee, Lexin Li, Tao Lu, Peter Marcy, Yonada Munoz, Juhyun Park, Debashis Paul, Jie Peng, Simon Preston, Sofia Olhede, Laura Sangalli, Damla Senturk, Jim Sethna, Valentina Staneva, Wenwen Tao, Michael Wierzbicki, Darren Wilkinson, Andy Wood, Donglin Zeng, Jay Wang, Naisyin Wang, Xiaohui Wang, Hulin Wu, Yuefeng Wu, Nuen Tsang Yang, Tingting Zhang, Xiaoke Zhang, Ji Zhu, Jian Zou

4.2 Topics

The field of statistical inference for nonlinear dynamic systems associated with observational data is a relatively new area of statistical research, although the fundamental ideas can be found in time-domain time series analysis, econometrics, pharmacological modeling and a few other topics with considerable histories. There are no general monographs on the modeling of data over continuous time that target specifically the statistical community.

Dynamic systems pose the wider challenge of statistical methodology for mathematical models expressed as systems of equations that do not admit an analytical solution. Classical approaches have involved iterative numerical approximation of solutions, but new methods are being proposed that use the equations themselves in regularization penalties in inverse problems, and consequently do not require repeated attempts to approximate exact solutions to these equations. Rather, these methods rely on relaxation strategies that have proven invaluable in other areas of numerical analysis.

Wider interest in the topic has been stimulated by new methods for parameter estimation and inference, the modeling of genomic and proteomic data, dynamic systems modeling of the spread and time course of disease, the use of stochastic dynamics beyond financial analysis, developments in functional data analysis, and the explosive growth in real-time data monitoring over space, time, and other continua and manifolds. This working group brought together several groups of statisticians who have separately developed methods from quite different perspectives, and several researchers interested in learning more about the problems in the area and how to get involved.

Because of these new developments, the working group was an invaluable opportunity for the consolidation of efforts, discussion of methodological and computational issues, and the exchange of perspectives. The working group therefore spent some considerable time reviewing literature and approaches and identifying new areas of potential interest. In particular, sources of interest were:

1. The distinction between deterministic and stochastic models of system dynamics and differences between them with respect to methodological approaches and fundamental conceptualization of modeling and inference problems.

2. Dealing with lack of fit due to simplifications or partial understanding inherent in any dynamic model for real world data.

3. The question of what is desirable to fit – data as observed, or (what is more frequently of interest to applications) qualitative features (cycles, periods, chaos) of the system dynamics.

4. The identifiability of parameters in dynamic systems models, and the problem of designing experiments to improve parameter estimation.

5. The design of software. This is particularly challenging in nonlinear dynamics: the models employed are non-linear and generally unique to each problem, this is coupled to models of

observation processes that are similarly problem-specific and these models need to be easily incorporated into computationally complex methods as well as being easily modified. Additionally, all current computational methods can suffer from failures, either of optimization or of stochastic sampling and useful diagnostic processes for these failures need to be developed.

6. The use of partial differential equations in spatial and spatial-temporal data analysis to model variation over space and time.

7. Ways of facilitating the entry of new researchers into this area, including books, collections of applications, workshops, on-line seminars and so on. The inclusion of nine lectures of 2 hours each on dynamical systems modeling in the second semester graduate seminar at SAMSI during the project was especially valuable as a stimulant to the consideration of methods for dissemination.

8. General methods for parameter estimation and inference for both continuous and discrete time dynamic systems.

9. A wide range of specific applications, including systems biology, spread of disease, viral dynamics, clinical trials, population dynamics in mathematical biology, glaciology, dynamic systems with covariates, chemical reactions, and transport models.

10. Ways of avoiding the heavy computational overhead and long development times typical of application of Markov Chain Monte Carlo methods for Bayesian analysis.

11. Parallel computing strategies for dynamic systems

12. A library of real-world datasets and models for testing software and methodology

The working group focussed on providing introductions to these problems and discussions of varying approaches to dealing with them within different modeling frameworks. In addition, several areas of statistical inference have been identified as future areas of research:

- Diagnostics for model lack of fit.
- Adaptive experimental design for nonlinear dynamics.
- Incorporating random effects into repeated dynamic systems.
- Optimal choice of qualitative features to assess both in terms of information obtained and robustness to model choices.
- The extension of methods for ordinary differential equations to partial differential equations. In this context, an example data set involving ice-melt on glaciers has been identified as a relevant and approachable problem.

4.3 Achievements, Working Papers and Publications

Achievements

Jim Ramsay translated the CollocInfer package in R into Matlab in January, and did considerable additional work on the CollocInfer manual and on various test problems. Giles Hooker and Jim Ramsay are continuing to collaborate on the development of this software for parameter estimation and inference for dynamic systems.

Published/Accepted

1. Cao, J., J. Cai and L. Wang (2011) Estimating Curves and Derivatives with Parametric Penalized Spline Smoothing, Accepted by Statistics and Computing.
2. Hooker, G., S. P. Ellner, L. Roditi and D. J. D. Earn (2011) Parameterizing State-space Models for Infectious Disease Dynamics by Generalized Profiling: Measles in Ontario, Journal of the Royal Society Interface, 8:961-975.
3. Wilkinson, D. J. (2011) Stochastic Modeling for Systems Biology, second edition, Boca Raton, Florida: Chapman and Hall/CRC Press, in press.
4. Wilkinson, D. J. (2011) Stochastic dynamical systems, in Handbook of Statistical Systems Biology, M.P.H. Stumpf, M. Girolami, D.J. Balding (eds), Wiley, in press.
5. Wu, R., J. Cao, Z. Huang, Z. Wang, J. Gai and E. Vallejos (2011) Systems mapping: how to improve the genetic mapping of complex traits through design principles of biological systems, BMC Systems Biology 5:84, 1-24.

Preprints/Technical Reports

1. Golightly, A., Wilkinson, D. J. (2011) Bayesian Parameter Inference for Stochastic Biochemical Network Models using Particle MCMC, in submission to Interface Focus.
2. Luo, W., J. Cao, M. Gallagher and J. Wiles (2011) Estimating the Intensity of Ward Admission and its Effect on Emergency Department Access Block, Submitted
3. Ratmann, O., P. Pudlo, S. Richardson and C. Robert, Monte Carlo Algorithms for Model Assessment via Conflicting Summaries, submitted.
4. Hooker, G. and S. P. Ellner, (2011) On Forwards Prediction Error, Technical Report BU-1679-M, Department of Biological Statistics and Computational Biology, Cornell University.
5. Hooker, G., J. O. Ramsay and L. Xiao, (2011) CollocInfer: An R Library for Collocation Inference for Continuous and DiscreteTime Dynamic Systems, R library and manual.

Working Papers/In Progress

1. Ramsay, J. O. (2011) A Functional Estimate of a Functional Variance-Covariance Matrix and its Inverse. *Paper in preparation.*
2. Sangalli, L., Ramsay, J.O. and Ramsay, T.O. (2011) Spatial Spline Regression Models. *Paper in preparation.*
3. Armagan, A., D. Dunson and J. Lee, Posterior consistency of Bayesian regression model for high-dimensional data.
4. Chkrebtii, O. and J. Cao, Generalized Additive Modeling of Spatio-Temporal Trends in Salmon Productivity
5. Cai, J., and J. Cao, Estimating Damage Accumulation Dynamic Models Hierarchical Bayesian Methods for Parameter Estimation in Dynamic Duration of Load Models
6. Dass, S., J. Lee, and K. Lee, Fast Computation for Regression Models with Ordinary Differential Equations.

7. Hooker, G., Rogers, B., Lin, K. and Ng, T., Control Theory and Optimal Adaptive Experiments in Nonlinear Stochastic Models

8. Ramsay, J. O. A Functional Estimate of a Functional Variance-Covariance Matrix and its Inverse

9. Sangalli, L., Ramsay, J.O. and Ramsay, T.O. Spatial Spline Regression Models Thorbergson, L. and Hooker, G., Optimal Adaptive Experimental Design in Hidden Markov Models

10. Wu, Y. and S. Ghosal, Convergence Rates of Multivariate Density Estimation by Dirichlet Mixture Priors

11. Wu, Y. and G. Hooker, Generalized Profiling, Stochastic Differential Equations and Higher-Order Stochastic Runge Kutta Schemes

5 Report for the Working Groups on Shapes and Manifolds

Leaders: Ian Dryden (University of South Carolina), Ezra Millar (Duke University), Victor Patrangenaru (Florida State University), John Kent (university of Leeds), Anuj Srivastava (Florida State University), Stephan Huckermann (University of Göttingen), Ross Whitaker (University of Utah).

5.1 Introduction

The Shapes and Manifolds theme involved three Working Groups:

1. Data analysis on sample spaces with a manifold stratification. Leaders: Ezra Miller, Victor Patrangenaru.

2. Metrics on shape spaces. Leaders: John Kent, Anuj Srivastava.

3. Geometric correspondence. Leaders: Stephan Huckemann, Ross Whitaker.

Individual reports from the working group leaders follow, but collectively many important issues were addressed during the program. There was a strong emphasis on sample spaces that have lower dimensional strata or sub-spaces, that themselves are manifolds. Particular examples include: metric phylogenetic trees on a fixed set of n taxa, where the space can be viewed as having structure like an open book, with pages attached to a book spine; covariance matrices with some equal eigenvalues; and the shapes of three dimensional landmarks, where lower dimensional shapes such as collinear points are also on a manifold. Some unexpected central limit theorem results were obtained, including various types of 'sticky results' where the Fréchet mean remains in a stratum with probability 1, as in some tree data examples. This contrasts with the more usual situation where Fréchet means are almost surely on the highest dimensional manifold, e.g. in the 3D shape case for certain Procrustes distances. Careful characterization of these issues was made by the Manifold Stratification Working Group, and this was an example of a project crossing both the Trees and Manifolds themes.

Excellent progress was made on the registration of functional data using curve shape analysis methods. This work initially arose out of a series of lectures and discussions in the Metrics working group. By adapting the Fisher-Rao metric based procedures for curves, some promising function registration methods were developed. This topic straddled both the Functional Data Analysis and Manifolds themes.

Careful examination of the topology and geometry of projective shape space was carried out by various members of the three groups. A key aspect of the work was to distinguish between an orientated versus an unorientated camera, and an axial versus directional camera. Using a particular representation called Tyler standardization the group was able to characterize all special cases. An alternative method of representing projective shapes based on projective frames was also investigated. These two approaches pave the way for practical methodologies for the statistical analysis of projective shapes from digital camera images.

A sub-group of the Metrics Group developed methodology for a type of spline on manifolds. By making use of the Palais metric it is possible to obtain an expression for the gradient of the associated objective function, which leads to a practical fitting algorithm. Details were worked out for various manifolds, including the space of rotations in 3D, Kendall's shape space and symmetric positive definite matrices with determinant 1. Also a new metric for the space of symmetric positive definite matrices was explored.

Dimension reduction was explored through influential landmarks. A sub-group of the Geometric Correspondence group developed a method for dimension reduction of shapes, by considering nested shape spaces where the most important landmarks are selected at each level. A large number of additional projects were also carried out on the general theme of Shapes and Manifolds, undertaken jointly by group members visiting SAMSI or resulting from discussions between groups members at SAMSI. A complete list is given in the individual working group reports.

All three Working Groups held regular seminar series, and extensive discussions. Coherent groups of researchers continue to work on topics of interest in this theme, and there were two sessions at the JSM 2011 in Miami Beach, Florida, which contained summaries of SAMSI material from the Shapes and Manifolds theme. A future workshop will be held in May 2012 at the Mathematical Biosciences Institute in Ohio, and the organizers of this workshop were all key participants in the Shapes and Manifolds theme.

The following subsections are reports from the working group leaders on the activities of the three manifolds working groups.

5.2 Report of Working Group: Data Analysis on Sample Spaces With Manifold Stratification

Leaders: Ezra Miller and Vic Patrangenaru

5.2.1 Overview and general research questions

At the opening workshop in Fall 2010, out of conversations between Vic Patrangenaru, Ezra Miller, and Stephan Huckemann following their presentations in 2-minute madness, it became clear that the objects the speakers in the meeting were studying all shared the desirable property of being represented as points on certain metric spaces that admit a *topological stratification*: a decomposition as a disjoint union of manifolds satisfying substantive tameness requirements (roughly: the singularities are "locally constant" on each stratum). Arguably, all object data may be represented as points on such spaces; therefore they should play a key role in twenty-first century Statistics.

Stratified sample spaces include real algebraic varieties (shape spaces or configuration spaces), polyhedral complexes (such as phylogenetic tree space), spaces of positive semidefinite matrices (from diffusion tensor imaging or any application producing distributions of covariance matrices). Our study involved a mix of geometry, probability, combinatorics, and computation, along with more usual statistical methods. The emerging field of *(geometrically) stratified statistics* elucidates phenomena in nonparametric multivariate statistical data analysis that arise when the sample space is singular. For example, tangential approximation does not yield standard statistics on Euclidean vector spaces, because singular spaces are not locally approximated by linear spaces. Therefore more geometric substitutes or analogues are required for notions such as mean, variance, principal component analysis, multi-dimensional scaling, and nonparametric bootstrap.

Historically, early results on nonparametric data analysis on Euclidean spaces, spheres, projective spaces, and orthogonal groups were pioneered by Daniel Bernoulli, deMoivre, Gauss, Cramer, Fréchet, G. Watson, David Kendall, Rudy Beran, Nick Fisher, Peter Hall, Andy Wood, John Kent, Huiling Le, Harrie Hendriks, Frits Ruymgaart, Peter Kim, Ted Chang, Herbert Ziezold, and other researchers. Data analysis on submanifolds of numerical spaces was first considered by Harrie Hendriks and Zinoviy Landsman. The first approach to nonparametric data analysis on abstract manifolds was due to Harrie Hendriks, Vic Patrangenaru, and Rabi Bhattacharya, and, later, to Peter Kim, Bruno Pelletier, Abhishek Bhattacharya, Ian Dryden, and others. On the other hand, nonparametric methods for data analysis on certain sample spaces with singularities (tree spaces or shape spaces) were developed by Susan Holmes, Steve Marron, Scott Provan, Megan Owen, Ezra Miller, Axel Munk, Stephan Huckemann, and others. These developments led us to the natural idea of developing a general approach to nonparametric data analysis on topologically stratified spaces, to extend the above methodologies for object data analysis, and to potentially address functional data analysis as well as analysis of shapes of curves or surfaces.

Collaboration between mathematicians, applied mathematicians, biostatisticians, statisticians, and computer scientists became a major theme (and goal) of the WG.

5.2.2 Key ideas and developments

1. **Statistics on negatively curved stratified spaces**

   The most fundamental phenomenon distinguishing singular sample spaces from smooth ones is that of *sticky means*: when a space is "negatively curved" near a singularity, large-sample empirical Fréchet sample means of arbitrary distributions can rest exactly at the Fréchet population mean, rather than asymptotically approaching the population mean as it does classically. Many of the WG discussions centered around quantifying this phenomenon and defining concepts so as to make it precise and gauge its generality. In particular, the WG investigated the shapes of central limit theorems in this context, particularly on spaces related to tree space (see the next subsection).

2. **Statistics on tree space**

   In the context of the previous paragraph, the motivating example for the WG of a negatively curved ("CAT(0)", or "globally nonpositively curved") singular space was the space T_n of metric phylogenetic trees on a fixed set of n taxa (Billera, Holmes, and Vogtmann 2001). It is a polyhedral complex whose number of facets is roughly factorial in the number of leaves. An algorithm to compute shortest paths in T_n efficiently was published shortly before the WG started (Owen and Provan 2011), and algorithms for centroid computation were in progress (Miller, Owen, and Provan 2011; Holmes 2011). Evolutionary biologists study distributions on T_n arising from gene tree or species tree reconstruction using genetic, morphological, or proteomic data. Notions of statistics on T_n are not new (see the survey Holmes (2003): *Statistics for phylogenetic trees*), but many issues remain open. The WG aimed to develop stratified methods for statistics and visualization specific to tree space and applications in evolutionary biology and medical imaging. Members of the WG carried out a number of computational experiments on statistical behavior of Fréchet means and other phenomena on polyhedral spaces. The tree space aspect of this WG overlapped substantially with the WG on Trees.

3. **Probability and statistics on positively curved stratified spaces**

Configurations of points can be viewed as discrete approximations to shapes in two or more dimensions. The set of equivalence classes of point configurations under group operations such as isometries, direct similarities, affine transformations, projective transformations, or other non-linear transformations is a *shape space*. Such spaces are, by definition, quotients of vector spaces modulo Lie group actions. Kendall shape spaces in dimension 3 or higher are important examples of stratified sample spaces with positive curvature. The role of singularities and stratifications in the positively curved context are markedly different from negatively curved situations, because Fréchet means tend to run away from the singularities here. In addition, the possibility of a nontrivial cut locus complicates the probability and statistics, since in this case there are no known conditions for the existence of a Fréchet mean.

4. **Probability and statistics on other stratified spaces**

   The WG spent substantial time discussing affine shape and projective shape, with applications in (for example) pattern recognition and machine vision in mind. Projective shape is widely accepted as the most appropriate notion of shape by computer scientists and others dealing with machine vision (Hartley and Zisserman). For this reason, two approaches were discussed by the WG for dealing with a statistical analysis of projective shapes in general position: one extends Patrangenaru's projective frame approach (J. Multivariate Anal. 2010), in which a specific base frame is fixed for the computation, and the other is based on a representation of projective shape by "Tyler standardization" (Tyler, Annals of Statistics, 1987; Biometrika 1987), in which every shape is brought to a projectively equivalent position that is optimal in a particular sense.

   The WG also discussed spaces of positive semidefinite matrices (covariance matrices), which play an important role in Diffusion Tensor Imaging. Again, two points of view were discussed: as a convex set in the space of matrices, and stratified according to which subsets of the eigenvalues are equal.

5. **Computational Issues**

   What type of data analysis is computationally faster on manifolds—and by extension on stratified spaces—was another issue discussed in the WG. Conversations on this topic compared and contrasted intrinsic and extrinsic analysis on shape and tree spaces as well as base-frame and Tyler-standardized approaches to image analysis. In the context of tree space, the WG had many discussions about the efficiency of various iterative algorithms to compute or approximate means.

6. **PCA on manifolds**

   The WG studied variants of PCA on manifolds, including the "forward approach" due to Fletcher et al. and the "backward approach" due to Huckeman et al. These two approaches both attempt to construct optimal nested families of submanifolds, generalizing the way PCA constructs a chain of vector subspaces by using increasing numbers of eigenvalues. The difference is that the forward approach constructs the nested submanifolds from low dimension to high dimension (essentially by PCA on the tangent space to the barycenter), while the backward approach starts with high dimension and proceeds downward by non-PCA optimization methods. Generalizing to singular stratified situations requires substantial ingenuity; it is the subject of some research projects generated by the WG.

5.2.3 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 40 WG members, but a few members did not actively participate after signing up to inspect the details of the WG.) Participation ranged from nearly always absent to nearly always present. The WG administrators were Vic Patrangenaru (Florida State U) and Ezra Miller (Duke). The other participants were Karthik Bharath, Marius Buibas, Michael Crane, Arturo Donate, Ian Dryden, Leif Ellingson, David Groisser, Harrie Hendriks, Stephan Huckemann, Sungkyu Jung, John Kent, Peter Kim, Yongdai Kim, Huiling Le, Xiuwen Liu, Steve Marron, John Moriarty, Daniel Osborne, Megan Owen, Frits Ruymgaart, Armin Schwartzman, Gabe Silva, Ross Whitaker, Andy Wood, Nuen Tsang Yang, Hongtu Zhu.

5.2.4 Activities

We had a two-hour meeting every week from mid-September 2010 until the beginning of June 2011, 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. Occasionally the meetings were joint with the Tree WG. During the last part of the program, the Geometric Correspondence WG joined our WG, and a few broadcast meetings in Spring 2011 were focused on related Geometric Correspondence topics.

The research in the WG was highlighted in the presentations by Patrangenaru (including contributions by many WG members), by Dryden (including contributions by Kent), and by Jung and Marron (joint work with Huckemann and Hotz) at the June 2011 SAMSI Transition Workshop.

Two WG members (Marron and Miller) presented relevant WG results at the AOOD Meets Evolutionary Biology conference at SAMSI, April 30–May 2, 2011. At least seven papers were presented by WG members at the 2011 Joint Statistical Meetings in Miami Beach, FL. Four WG members (Hendriks, Huckemann, Miller, Patrangenaru) attended a conference on “Nonparametrics and Geometry”, in Prague, Czech Republic. Three of them presented research connected to the WG at the conference.

A workshop proposal on the topics of the WG, written by Ezra Miller and submitted to the Mathematical Biosciences Institute (MBI), was funded to take place on 21–25 May 2011, with the title “Workshop on statistics, geometry, and combinatorics on stratified spaces arising from biological problems”. The organizers are Miller along with Huckemann, Le, Owen, and Patrangenaru. At least 17 of the WG participants, including all of the key players, will attend that workshop, but it is important to note that two dozen additional prominent researchers in related areas have accepted invitations to participate or speak.

5.2.5 Working Group research output

The group has been extremely productive and most of the resulting papers in preparation or published are listed in the following annotated references.

1. Thomas Hotz, Stephan Huckemann, Huiling Le, Steve Marron, Jonathan Mattingly, Ezra Miller, James Nolen, Megan Owen, Victor Patrangenaru, Sean Skwerer (2011), Sticky central limit theorems on open books, *preprint*, (with some subset of the above as authors).

2. Stephan Huckemann, Jonathan Mattingly, Ezra Miller, James Nolen (2011), Central limit theorems in codimension 1 on nonpositively curved stratified spaces, *work in progress.*

3. Steve Marron, Ipek Oguz, Sean Skwerer (2011), Smoothing in phylogenetic tree space using simple iterated pairwise geodesics, *work in progress.*

4. Stephan Huckemann, Steve Marron, Ezra Miller, Yolanda Munoz, Megan Owen, Victor Patrangenaru, Sean Skwerer (2011), Multidimensional scaling (MDS) with curved targets, *work in progress.*

5. Megan Owen, Sean Skwerer (2011), Fréchet means in tree space, *work in progress.*

6. Rabi Bhattacharya, Leif Ellingson, Xiuwen Liu, Vic Patrangenaru, Michael Crane (2011), Extrinsic analysis on manifolds is computationally faster than intrinsic analysis, with applications to quality control by machine vision, *to appear in Appl. Stochastic Models in Business and Industry.*

7. Steve Marron, Ezra Miller, Megan Owen, Scott Provan, Sean Skwerer (2011), Towards PCA on tree spaces, *work in progress.*

8. Thomas Hotz, Stephan Huckemann (2011), Intrinsic means on the circle: uniqueness, locus, and asymptotics. *http://arxiv.org/abs/1108.2141 [stat.ME] [math.PR]*

9. Leif Ellingson, Frits Ruymgaart, Vic Patrangenaru (2011), Nonparametric estimation for extrinsic mean shapes of planar contours,*revision submitted to Ann. Statist.*

10. John Kent, Kanti Mardia (2011), The geometric approach to projective shape and the cross ratio,*paper in preparation.*

11. John Kent, Thomas Hotz, Stephan Huckemann, and Ezra Miller (2011), The geometry and topology of projective shape spaces, *preprint.*

12. Thomas Hotz, Stephan Huckemann, Huiling Le, Vic Patrangenaru (2011), Hyperbolic data analysis, *work in progress.*

13. Leif Ellingson, Vic Patrangenaru, Sean Skwerer (2011), Computational methods for extrinsic mean on phylogenetic tree spaces, *work in progress.*

14. Leif Ellingson, Vic Patrangenaru (2011), Extrinsic PCA on manifolds, *work in progress.*

15. Leif Ellingson, David Groisser, Daniel Osborne, Vic Patrangenaru, Armin Schwartzman (2011), Data analysis on spaces of positive definite matrices with an application to dyslexia detection, *preprint.*

16. Harrie Hendriks (2011), Two sample problem for mean location, *http://nonparam11.karlin.mff.cuni. cz/bookabs20110719.pdf*

17. Ezra Miller, David Houle, Paul Bendich (2011), Quantifying shape differences in fruit fly wing morphology using persistent homology, *in progress.*

18. Marius Buibas, Michael Crane, Leif Ellingson, Vic Patrangenaru (2011), A projective frame based shape analysis of a rigid scene from noncalibrated digital camera imaging outputs, *to appear in Proc. of JSM, 2011, Miami, FL.*

19. Sungkyu Jung, Stephan Huckemann, J. S. Marron, Thomas Hotz (2011), Principal nested shape spaces and an application to reduction of number of landmarks, *work in progress, http://www.samsi.info/workshop/aood-transition-workshop-june-9-11-2011.*

20. Daniel Osborne, Victor Patrangenaru, Xiuwen Liu, Hillary Thompson (2011), 3D size-and-reflection shape analysis for planning reconstructive surgery of the skull, *to appear in Proc. of JSM, 2011, Miami, FL.*

5.3 Report on Working Group: Metrics on Shape Spaces

Leaders: John Kent and Anuj Srivastava

"Shape" consists of the information in a geometric object that is invariant under a group of transformations. For objects consisting of a set of labeled points in $\mathbb{R}^d$ (typically $d = 1, 2$ or 3), these groups can include some or all of the following operations: translation, scaling, rotation, reflection and, more generally, projective transformations. For continuous objects consisting of unlabeled points such as curves and surfaces, changes to the parameterization of the object can also be included (e.g. time-warping of curves).

It is also possible to think about shape more generally, where in addition to the location of each point of an object, there is also extra information recorded at each point. Examples include a direction or a positive definite matrix. The space of shapes forms a manifold-like structure and it is important to be able to quantify differences in shape. Procrustes methods generalize Euclidean distance. For continuous objects, nonlinear deformations can also be included.

The overall goal of the working group was to explore the implications of the choice of metric in various shape spaces. A metric often induces a Riemannian structure, which in turn leads to geodesics and parallel transport. The reason for our interest in metrics is that different metrics can highlight different features of shape differences.

5.3.1 Key ideas and developments

The participants in the group had a wide variety of backgrounds ranging from engineering and image analysis to pure and applied mathematics and statistics. We discovered that many of us were tackling similar problems from different points of view, and the working group provided an opportunity for us to deepen our understanding and appreciate connections to other areas, and to build new collaborations. The unifying theme of the working group is perhaps best described as new methods to deal with "shape", defined very generally.

Several themes emerged through the discussions: the importance of key concepts in differential geometry (Riemannian metrics exponential function, parallel transport), time-warping (Fisher-Rao metric), fitting curves (especially growth curves with more general links to functional data analysis), methods for positive definite matrices, a deeper understanding of projective shape and advances in the related field of directional data analysis.

Application areas include human activity modeling using image analysis, object identification in image analysis, growth curves, handwritten signatures, protein shape, computer vision, diffusion weighted magnetic resonance imaging, neuroscience spike trains, and gene expression signals. Here is a summary of the main achievements, where the numbering of the references refers to Section 5.3.4.

1. Functional data

   Ideas originally developed for the shapes of curves (dealing with the deformation of time using the Fisher-Rao metric) have been adapted to align curves in functional data analysis. A good example is growth curves, where growth spurts may occur at different times in different individuals. [4]

   Splines are a tool for fitting nonparametrically smooth functions of time. Classical spline theory provides exact methods for fitting functions with values in Euclidean spaces; simple approximate methods have also been developed for functions with values in simple manifolds such as spheres. New work gives more sophisticated fitting methods on more general manifolds. Applications include modeling a video sequence of a dancing figure. [1,2]

2. Projective geometry

   A projective shape consists of the features of a configuration of points that are invariant under different camera views. Classically projective shape has been studied using projective invariants such as the cross ratio. However, projective invariants do not possess a natural metric structure to enable the comparisons of different projective shapes. A standardized representation of the configuration has recently been proposed to facilitate such comparisons, with links to similarity shape analysis. As a result of the Samsi program, a much deeper understanding of the topology and geometry of this standardization is now available, especially at singularities. [5,6] [This work also fits into the theme of the working group on Data Analysis on Sample Spaces with a Manifold Stratification.]

3. Directional data analysis

   Classically this subject is concerned with the directions of points on a circle or sphere, but there are natural extensions to other objects involving an orientation, such as the eigenvalues and eigenvectors of a symmetric matrix. New nonparametric kernel density methods using mixtures of Bingham distributions have been developed, with applications to the use of diffusion weighted magnetic resonance imaging in the brain to model white matter fiber orientation. [3]

   A protein consists of a sequence of amino acids in three-dimensional space and its shape can be described in terms of a collection of angles on the circle. New methods based on kernel density estimation and on mixture modeling have been developed in this setting to give a deeper understanding of protein structure. [10,11]. This work is somewhat tangential to the main interests of the working group, but was stimulated during Mardia's visit to Samsi by his contacts with the Richardson Lab (Professors David and Jane Richardson) at Duke University.

4. Feature selection in images

   Differential invariants are a version of projective invariants which have been developed to pick out features of curves in images which are invariant under different camera views. [7]

   Applications such as human activity analysis can be viewed as curves in a similarity shape space (a manifold). Using ideas such as parallel transport, such paths can be transformed into curves in Euclidean space for the purposes of statistical identification, comparison and dimension reduction. [8,9]

5.3.2 Participants

John Kent was in charge of organizing the group's activities, with Anuj Srivastava also being a joint leader. The following list includes everyone who showed an interest in the working group, with participation ranging from slight to regular: James Damon, Ian Dryden, Jinjiang He , Stephan Huckemann, Sungkyu Jung, Hachem Kadri, John Kent, Irina Kogan, Hamid Krim, Sebastian Kurtek, Huiling Le, Kanti Mardia, Jeffrey Morris, Hans Mueller, Megan Owen, Victor Panaretos, Hailin Sang, Christof Seiler, Anuj Srivastava, Valentina Staneva, Jonathan Taylor, Alain Trouvé, Jane-Ling Wang, Andy Wood, Ross Whittaker, Christine Xu, Nuen Tsang Yang, Laurent Younes, Hongtu Zhu.

5.3.3 Subgroups and their topics, activities

We had a two-hour meeting every week, from mid-September until the end of November 2010, with occasional, largely remote, meetings in the period Jan - March 2011. Typically one person would make a presentation about current achievements or work in progress. This would sometimes inspire smaller groups to work on problems in more detail, with summary results often reported back to the working group. Some talks and discussions:

- Anuj Srivastava: Shapes of elastic curves.
- Hans Mueller: Function registration.
- Sebastian Kurtek: Riemannian framework for function registration.
- Hamid Krim: Squigraph modeling of shapes
- Irina Kogan: Geometric transformation, invariance, curve matching.
- John Kent: Basis expansion of shape. Projective shapes.
- Valentina Staneva: Diffeomorphisms and RKHS.
- Ian Dryden: data-based distance metrics.

A summary of the work from the Metrics groups was presented by Ian Dryden at the Transition Workshop, as well as talks by Vic Patrangenaru, Sungkyu Jung, Sebastian Kurtek and Jingyong Su. Also, work was presented at two sessions at JSM, 2011.

5.3.4 List of papers (preprints) of work done at SAMSI or inspired by SAMSI

Several research papers have been substantially influenced by discussions and interactions at Samsi.

1. J. Su, I. L. Dryden, E. Klassen, H. Le and A. Srivastava (2011), Fitting Smoothing Splines to Time-Indexed, Noisy Points on Nonlinear Manifolds. *Submitted to Journal of Image and Vision Computing.*
2. J. Su, I. L. Dryden, E. Klassen, H. Le and A. Srivastava, A new metric for Symmetric Positive Definite Matrices. *Technical report in preparation.*
3. Dryden, I.L. and Olhede, S. (2011), A kernel density estimator for the orientation distribution function in diffusion weighted imaging, using Bingham Distributions. *Technical report in preparation.*
4. A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron, Registration of Functional Data Using Fisher-Rao Metric. *Technical report.*
5. J T Kent and K V Mardia, The geometric approach to projective shape and the cross ratio. *Technical report in preparation.*
6. John Kent, Thomas Hotz, Stephan Huckemann and Ezra Miller, The geometry and topology of projective shape spaces. *Technical report in preparation.*
7. Burdis J. M., Kogan, I.A., Object-image correspondence for curves under finite and affine cameras.
   Technical Report. http://arxiv.org/abs/1004.0393
8. Sheng Yi, Hamid Krim, and Larry K. Norris, Human Activity as a Manifold Valued Random Process. *IEEE Transactions on Image Processing*

9. Sheng Yi, Hamid Krim, Larry K. Norris, A Invertible Dimension Reduction of Curves on a Manifold.
Technical report. arXiv:submit/0291586 [cs.CV] 29 Jul 2011

10. Charles C. Taylor, Kanti V. Mardia, Marco Di Marzio, Agnese Panzera, Validating protein structure using kernel density estimates. *Submitted to Statistical Applications in Genetics and Molecular Biology.*

11. K. V. Mardia, J. T. Kent, Z Zheng, C. C. Taylor and T. Hamelryck, Mixtures of concentrated sine distributions with applications to bioinformatics. *Technical report in preparation.*

5.4 Report for Workgroup: Geometric Correspondence

Leaders: Stephan Huckemann and Ross Whitaker

5.4.1 Participants and Meetings

Stephan Huckemann (administrator), Ross Whitaker (administrator)

Andy Wood, Anuj Srivastava, Christine Xu, Christof Seiler, Clarisa Williams, Cong Xu, Daniel Gervini, Gosh Debashis, Ezra Miller, Giseon Heo, Hachem Kadri, Haonan Wang, Ian Dryden, James Damon, Jeffrey Morris, John Kent, John Moriarty, Kanti Mardia, Kaushik Ghosh, Leif Ellingson, Nuen Tsang Yang, Sean Skwerer, Serge Guillas, Steve Marron, Sue McDonald, Sungkyu Jung, Valentina Staneva, Victor Patrangenaru, Yongdai Kim

After spinoff the group met weekly (Thursdays 1–3 pm) with the bulk of participants having physically left SAMSI, joining over the webex interface. A period of lively discussions and formation of working subgroups followed in the months of fall. When finally also the administrators physically left SAMSI it was decided to continue on working subgroup level while the main group meetings were joined with group weekly meetings of the working group "Data Analysis on Sample Spaces with a Manifold Stratification".

5.4.2 Thematic Content

After singling out key issues of interest, a series of talks followed which ultimately led to the formation of five working subgroups. The working subgroups began to have regular meetings independent of the main group meetings which were eventually joined with another workgroup. The working subgroups are currently still very active and meet on a regular basis, often weekly.

1. **Identification of Key Issues**

   Our first meetings were dominated by identifying key issues occurring in the analysis of geometrical objects in view of geometric correspondence. Four main topics emerged. The first two reflecting different means of representations, the third an overall underlying statistical approach and the fourth emphasizing the connection of research with driving problems in applications.

2. **Alignment of Functions**

   In this context it became clear that future work would confront the difficult issue of "alignment" without "knots". Discussions focussed around more strongly exploiting the joint information given by ensembles of functions of interest. Also, the group got excited about an idea proposed by the group member Daniel Gervini going from the traditional "horizontal alignment" toward a "vertical alignment".

3. **Alignment of Shapes**

   It was agreed that all infinite dimensional shape representations should be dealt with in the above context. Here, we would concentrate on shape representations given by a finite number of landmarks. Key issues not having found a satisfactory solution to date consist in the alignment of "mismatching configurations". This comprises the problem of treating occlusions, topologcial differences, unknown labeling or even unknown positioning. It became very clear that ideas borrowing from statistics could be powerful and robust enough to tackle the enormous computational challenges involved. In this context the group grew very interested in previous work of the group leader Ross Whitaker, optimizing an information theoretic content as well as in the non-deterministic "pseudo landmarking" introduced by the group member Vic Patrangenaru. Also, based on the method of "principal nested spheres" recently developed by the group members Steve Marron and Sunkyu Jung, a potential for the indentification of the statistical importance of a single landmark was noted by the group leader Stephan Huckemann.

4. **Statistical Issues**

   The above topics let to intense discussions of statistical issues and their intertwining with geometrical ideas. Within a non-Euclidean geometry which is a natural surrounding in the context of the problems at hand, elementary statistical issues e.g. regression, avoiding overfitting, the role of bias and the use of fair "p"-values turn into non-trivial challenges. In the specific geometric context, identification of information can be obtained by marginalizing over the deformation versus residual variability. Here, it was suggested to investigate an iterative "minimax" algorithmic approach.

5. **Driven by Applications**

   Several group members, including Kanti Mardia, stressed the urgent neccessity to develop sophisticated tools to address current problem in the areas of protein matching and folding. Fairly recent approaches based on angular information lead to tori model spaces which turn out to be unsatisfactory e.g. for PCA like methods. It was expected that the discussion of the above collected issues provides for a methodology to link in a genuine 3D fashion deformaton with chemistry, possibly exploiting a matching of "blobs" between structures. Such methodology may also turn out to be successful for any other comparison between organisms or parts thereof, be it ensembles of comparatively smooth surfaces of biological shapes (bones, brains, faces, hippocampi, etc.) or branching structures occuring among others in medial skeletal modeling and phylogenetic trees.

   An important aspect was added by the group member Vic Patrangenaru, pointing out that all human visual information is not based on similarity shape but on projective shape. This led to a feed back to the "Alignment of Shape" subgroup to develop a "frame-free" formulation of projective shape. Motivated by applications, we also touched the issue of treatment of shape discontinuities, and the issue of alignment "under constraints", e.g. growth.

5.4.3 Series of Talks

After a series of rather informal discussions identifying key issues of interest, group meetings provided for ample space for extended talks and intense discussions. Until the end of November 2010 the following talks and follow up extended discussions were held by core group members from their individual expertise, touching the above key goals.

- Automated correspondences (Ross Whitaker),

- functional alignment (Daniel Gervini),
- curve matching landmarks (Leif Ellingson and Victor Patrangenaru),
- alignment within Gaussian processes (John Moriarty) and
- correspondence using random fields (Ian Dryden).

5.4.4 Working Subgroup Formation

Following these talks and discussions, certain collaborative work materialized in specific research projects. The corresponding subgroups began to meet regularly.

- **Measuring Landmark Importance for Planar Shapes** Planar landmark based Kendall shape space is a complex projective space. A space with less landmarks is a lower dimensional complex projective space that can be embedded in various ways. From a given data sample with k landmarks, in previously unpublished research, a method has been developed to compute the best fitting abstract shape space with $k-1$ landmarks and in this the best abstract shape space with $k-2$ landmarks, etc. The entire procedure until the shape space of triangles is reached, or even further, up to a single geodesic and a single point can be called a *nested backward shape space principal component analysis* (cf. Jung et al., 2011, submitted). One way of assessing landmark importance is to compare the total intrinsic variances in these abstract spaces with the total intrinsic variances of concrete spaces obtained from leaving out one and more specific landmarks. Algorithmic work has begun as well as some applications to Ian Dryden's digits 3 dataset (Dryden and Mardia, 1998, Wiley) have been performed. Further application are planned to leaf data (cf. Huckemann et al., 2010, IEEE PAMI) in order to automatically identify a minimal number of landmark placment for specific discrimination problems arising in forestry. First results have been presented at the AOOD transition workshop and will be submitted for publication shortly:

  Huckemann, S., Jung S., Marron, S.(2011) Principal Nested Shape Spaces and an Application to the Reduction of the Number of Landmarks.

- **Modeling Locus and Shape Variation by Partial Rotations**

  A goal in computational anatomy is to keep track of rigid body motions as well as of deformations of internal organs, one application being radiation therapy of the male prostate. In a novel approach, we are estimating few axes of rotations of parts of organs in order to nearly exhaustively model rigid body rotation, and rotations of parts of the organ, e.g. bending and pinching. Preliminary research based on modifications of *principial arc analysis* for spoke data on medial skeletons (cf. Jung et al., 2010, submitted) has shown to have a very promising potential. Estimating dominating axes of (internal) rotations may also have applications to protein folding.

  This currently very active research has recently begun to attract numerous new collaborators and resulted in a first publication in a conference proceeding

  Pizer, S., Jung, S., Goswami, D., Zhao, X., Chaudhuri, R., Damon, J., Huckemann, S. and Marron, S.J. (2011). Nested sphere statistics of skeletal models, *in Proc. Dagstuhl Workshop on Innovations for Shape Analysis.*

  This research will also be presented at the Birs workshop, "Geometry for Anatomy" at Banff in August 2011.

- **A Frame-Free Projective Shape Space**

  All shape spaces come typically as equivalence classes on a "top" space from which – in order to obtain a meaningful topology – a certain "singularity set" has to be removed. For similarity shapes this singularity set is rather obvious. For projective landmark based shapes, a typical choice has been to select a frame in general position and to prohibit any other landmark to coincide with any of the frame points (e.g. Mardia and Patrangenaru, 2005). Obviously a statistical analysis is dependent on the frame chosen. Furthermore, it turns out that the resulting shape space may be disconnected. The components may be connected when certain landmark correspondences are tolerated. Then, among others, the resulting topology has to be carefully analyzed. This research has resulted in a collaboration with two other working groups "Metrics on Shape Spaces" and "Data Analysis on Sample Spaces with a Manifold Stratification". In fact the problems touched are very deep and fundamental for any shape analysis building on any groups other than the similarity group conveying "shape equivalence". A first preprint is in its finalizing stage:

  Kent, J., Hotz, T., Huckemann, S. and Miller, E. (2011) The topology and geometry of projective shape spaces, *Preprint.*

- **Automatic Landmark Extraction for Planar Contours**

  For the infinite dimensional shape spaces of closed contours, to date only an intrinsic statistical analysis is available coming along with the typical and quite considerable numerical challenges to compute intrinsic quantities. Moreover, for the minimization process involved, there is no guarantee that a local minimum found actually corresponds to the desired global minimimum. In order to make an extrinsic statistical methodology available also for infinite dimensional spaces, with simple numerical procedures yielding guaranteed global minimizers, for the space of planar curves one can consider the limit $k \to \infty$ for the extrinsic Veronese-Whitney embedding of Kendall's planar shape space Σ_2^k with k landmarks. Additionally, a desired boundary correspondence can be obtained by placing the k landmarks according to a suitable probabilistic model. Ongoing research into this direction has resulted at this point to

  Ellingson, L., Patrangenaru, V. and Ruymgaart F. (2011) Automatic Landmark Extraction for Planar Contours, *Preprint submitted for publication.*

- **Generative Models for Registering Ensembles of Functions and Images**

  Suppose we have a set of functions f_i of which have sampled values f_{ij}. Underlying are unwarped signals $\widetilde{f}_i = f_i \circ T_i^{-1}$ with a warping T_i. We want to estimate the T_i from the data by setting up a *generative process*, that in the first instance we consider Gaussian

  $$P(T, f) = G\big(\Sigma_{\widetilde{f}}, \mu_{\widetilde{f}}\big) G\big(\Sigma_T, \mu_T\big)$$

  which suggests a EM-approach to identify hidden parameters. E.g., such an approach for images warps images not to each other but rather to a mean. Currently active researchers are Daniel Gervini and Ross Whitaker.

- **Minimax Correspondence for Shape Classification**

  Many applications deal with a specific classification problem. Here we present a generic method of addressing this issue, which is of interest for applications in forestry, assessing leaf variation of leaf shape over genotype, temperature gradient and leaf location in the tree (e.g. crown or breast hight). Previous research using landmark based shape spaces has shown that for powers of tests the "correct" number and placement of landmarks is of great importance. In this project for a specific classification problem given, an automated landmarking scheme is to be developed that places landmarks such that the intra-group variation is minimal while simultaneously the variation over the groups is maximal. This method is to be based on purely statistical features (Cates et al., 2006, Int Conf Med Image Comput Comput Assist Interv). Another potential application of this method lies in the early diagnosis of degenerative brain processes such as Alzheimer's desease. Research in this project is currently at a very conceptional level, both working out a stringent mathematical formulation and making feasible a numerical approach. Currently active researchers are Stephan Huckemann, Ross Whitaker, Martin Styner.

6 Report for the Working Group: Brain Imaging

Leaders: John Aston (*Chair*), Jeffrey Morris, Hans Müller, Haipeng Shen and Jane-Ling Wang. SAMSI Postdoc and Group Webmaster: Ci-Ren Jiang.

6.1 Summary

This working group started out specifically focused on the link between functional data analysis and brain imaging. However, during the SAMSI AOOD year, the group extended its focus to all statistical issues associated with brain imaging data sets. In particular three subgroups were formed to examine the aspects of Deconvolution and Design-Free analysis, Spatial and Temporal Modeling, and Hierarchical analysis in Brain Imaging (this last subgroup being also associated with the Hierarchical Modeling SAMSI working group). The group had regular weekly meetings and talks during the Fall semester and more focused working group meetings (including talks and increasingly presentations on current work) during the Spring semester. Throughout the year, there were very close connections with the Functional Data Analysis groups, in particular the WG on hierarchical functional data also developed an emphasis on brain imaging and the resulting papers could also be co-opted by this WG.

6.2 Participants

The working group consisted of 103 members who regularly received the email notifications and updates on the group activities, making it one of the largest SAMSI working groups. Not all of these were active members, but there was still considerable activity throughout the year.

Active Participants: Adrian Bowman, Alois Kneip, Armin Schwartzmann, Bill Schucany, DuBois Bowman, Giles Hooker, Fan Li, Hernando Ombao, Hongtu Zhu, Ian Dryden, Juhyun Park, Lexin Li, Nuen Tsang Yang, Phil Reiss, Richard Samworth, Seonjoo Lee, Sylvie Tchumtchoua, Tingting Zhang, Todd Ogden, Xia Wang. Many additional participants joined for particular talks or group meetings, as well as the six or so outside speakers, who gave presentations on their work to the group.

6.3 Activities

As mentioned, there were 3 subgroups

1. Deconvolution and Model Free Analysis in Brain Imaging - leader Haipeng Shen. This working group mainly examined the problem of estimating parameters in time series analysis of functional Magnetic Resonance Imaging (fMRI) data where part of the experimental design

was unknown. One particular focus was Hemodynamic Response Function (HRF) estimation. In usual fMRI analysis, the HRF is assumed to be known and convolved with the experimental design before linear modeling takes place. The working group examined under different modeling assumptions how best to account for the HRF when it was assumed to be unknown. This lead to several innovative techniques being developed during the year, as was the focus of 2 talks during the overall SAMSI AOOD transition workshop, with an additional talk being focused on deconvolution models in Positron Emission Tomography. A second area of interest was the converse problem of when the HRF was assumed known but where the experimental design was now unknown. This can be framed in a change point context and several talks and projects throughout the year investigated this area.

2. Spatial Temporal Subgroup - leader John Aston. This subgroup explored the issues of spatial and temporal correlations in brain imaging and particular as to how this might affect techniques such as functional data analysis, which are typically assumed to have independent errors. This subgroup especially benefitted from many speakers from outside areas talking about their experiences for different data and the subgroup then explored how these ideas might be more focused on imaging data. Particularly areas of research included the problems of massive data structures on covariance estimation, accounting for temporal or spatial correlation in principal component analysis and functional basis approaches for representing imaging data.

3. Hierarchical Modeling - leader Jeffrey Morris. This working group emphasised looking at how multi-trial or multi-subject data could be analysed efficiently through various approaches. One aspect of this working group that was of particular interest in relation to the other subgroups was the inclusion of Bayesian modeling principles as well as frequentist approaches into the framework. fMRI was again the primary focus of this group, but other modalities were also considered. As mentioned earlier, this working group was somewhat of a cross-over subgroup between two working groups, the brain imaging working group and the hierarchical modeling working group.

The working group also took an active role in some of the larger workshops that occurred during the year. Several of the working group leaders (Hans Mller and Jane-Ling Wang) were also very much involved in the longitudinal data analysis workshop (with many brain imaging group members being active participants in this workshop). In the final AOOD transition workshop, the Brain Imaging working group (particular Jane-Ling Wang) organised the final half-day session with five talks in all areas covered by the working group. In addition, there was a final panel discussion at the end of the AOOD workshop primarily focussing on how to make the statistical methodologies of groups such as the Brain Imaging group more readily available and accessible to those working in the field of Brain Imaging as practitioners.

A few examples of projects inspired and/or carried out by the Brain Imaging Working Group

- UCD team (in alphabetical order: O. Carmichael, J. He, H.-G. Müller and J.-L. Wang) - We study resting state FMRI data that were collected at UC Davis and develop functional correlation measures. The goal is to quantify spatial dependencies of the FMRI signals at various voxels, aiming at improved connectivity measures and subject classification. The work is still ongoing and benefitted from discussions of the SAMSI Brain Working group. Part of the work is performed while Mueller and Wang visited SAMSI in Fall 2010, so we would like to acknowledge SAMSI and the brain working group.

- Tingting Zhang - I participated in SAMSI 2010-11 Program on Analysis of Object Data, because my research interest of brain imaging data analysis falls within the applications areas emphasized by the program. Under its support, I was able to visit SAMSI in 2011 spring, and developed collaborative projects with Dr. Ahmad Hariri, Professor of Psychology and Neuroscience and Investigator in the Institute for Genome Sciences and Policy at Duke University. My collaborator Dr. Fan Li from Duke University and I were motivated by SAMSI working group discussions. We submitted an NSF grant on functional imaging genetic data analysis. In addition, we are going to finish a paper on nonparametric inference of functional magnetic resonance imaging (fMRI) data. In a short summary, we have greatly benefited from the opportunities presented by the SAMSI program

- Phil Reiss - My work on voxel-by-voxel nonparametric inference for samples of brain images (in collaboration with Lei Huang of NYU) was inspired by discussions at the September and November 2010 workshops on Analysis of Object Data. In particular, a conversation with Ciprian Crainiceanu at the opening workshop led to a visit to his working group at Johns Hopkins, which gave me an opportunity to present an early version of the work and to have further helpful discussions. The final meeting of the Brain Imaging working group provided me with a valuable opportunity to present two functional data analytic aspects of this project: (1) a procedure for clustering of voxelwise function estimates, and (2) a modified smoothing procedure that borrows strength across neighboring voxels. The feedback received during my presentation helped a great deal to fine-tune this work, which is currently being prepared for publication.

6.4 Papers/Preprints/Work in Preparation

1. Aston JAD and Kirch C. Detecting and estimating epidemic changes in dependent functional data, *submitted.*

2. Aston JAD and Kirch C. Estimation of the distribution of change-points with application to fMRI data, *in preparation.*

3. Bunea F, She Y Ombao H, Gongvatana W, Devlin K and Cohen R. (2011). Penalized Least Squares Regression Methods and Applications to Neuroimaging, *NeuroImage*, (55), 1519-1527.

4. Fiecas, M. and Ombao, H. (2011). The Generalized Shrinkage Estimator for the Analysis of Functional Connectivity of Brain Signals, *Annals of Appl Stat, in press.*

5. Gorrostieta C and Ombao H.. General Spectral Measures of Cross-Dependence in Multivariate Time Series, *JASA, under revision.*

6. Gorrostieta C, Ombao H, Bedard P and Sanes J.N. Investigating Stimulus-Induced Changes in Connectivity Using Mixed Effects Vector Autoregressive Models, *NeuroImage, under revision.*

7. Gorrostieta C, Ombao H, Rrado R, Patel S and Eskandar E. Coherence Analysis of Local Field Potentials, *Journal of Time Series Analysis, under review.*

8. Jiang C, Wang JL and Aston JAD. Functional PCA based Deconvolution for Position Emission Tomography, *in preparation.*

9. Kang H, Ombao H, Linkletter C, Long N and Badre D. Spatio-Spectral Mixed Effects Model for Functional Magnetic Resonance Imaging Data, *JASA, under revision.*

10. Lee S, Shen H, Truong Y, Lewis M, and Huang X (2011) Independent Component Analysis Involving Auto-correlated Sources with an Application to Functional Magnetic Resonance Imaging, *Journal of the American Statistical Association, accepted.*

11. Morris JS, Baladandauthapani V, Herrick RC, Sanna PP, and Gutstein HG (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomic data, *Annals of Applied Statistics, 5(2A), 894-923.*

12. Nam CFH, Aston JAD and Johansen AM, *Quantifying uncertainty in change points, in revision.*

13. Ogden, R. T., Zhao, Y., and Reiss, P. T. Wavelet-based functional principal component regression, *In preparation.*

14. Shen H, Tian S, and Huang J. A Two-Way Regularization Method for MEG Source Reconstruction, *under review at Annals of Applied Statistics.*

15. Shen H, Trong Y, and Lee S. Asymptotics of Colored Independent Component Analysis, *under review at Annals of Statistics.*

16. Shen H, Zhang L and Huang J. Two-Way Robust Functional Data Analysis, *to be submitted shortly.*

17. Shen H, Truong Y and Chen W. Hemodynamic Response Function Estimation, *in preparation.*

18. Shi, XY, Zhu, HT, Ibrahim JG, Styner M, Intrinsic regression models for median representation of subcortical structures.

19. Tchumtchoua, S., Dunson, D. B., and Morris, J. Online Variational Bayes Inference for High-Dimensional Correlated Data, *In preparation (to be submitted soon).*

20. Tchumtchoua S, A Heterogeneous Dynamic Structural Equation Model with Application to Brain Connectivity, *in preparation.*

21. Van Lunen D, Ombao H and Aston JAD. Online Detection Methods: A Model Selection Framework. *In preparation.*

22. Yuan, Y., Zhu, H.T., Lin, W. L., and Marron, J. S. Local polynomial regression for symmetric positive definitive matrices.

23. Yuan, Y., Zhu, H.T., Styner, M., J. H. Gilmore., and Marron, J. S. Varying coefficient model for modeling diffusion tensors along white matter bundles.

24. Zhao, Y., Ogden, R. T., and Reiss, P.T. Wavelet-based LASSO in functional linear regression, *Submitted manuscript.*

25. Zhao, Y., Bagiella, E., and Ogden, R. T. A functional approach to analysis of RR interval variability, *In preparation.*

26. Zhu, H., Brown, P. J., and Morris, J. S. Robust classification of functional and quantitative image data using functional mixed models, *Submitted.*

7 Report for the Working Group: Tree Structured Data Objects

Leader: J.S. Marron

7.1 Overview

Following the Opening Workshop, and follow-on discussion, three main approaches emerged:

1. Combinatorial: There were some discussions of this approach, but the main players seemed to move outside the SAMSI framework, perhaps mostly due to SAMSI restrictions on meeting times.
2. Phylogenetic Trees: This group continued very actively through the entire year, and there were very interesting interactions with the Stratified Manifolds Working Group. Half of the meeting time each week was devoted to this effort.
3. Dyck Path: This group was also very active throughout the year, and occupied the other half of the meeting time.

In both of the latter two groups, activities centered around weekly discussions. In each case, there were both presentations of related work by mostly irregular participants, and discussion of on-going work led by regular participants. Much of the on-going work was done by a single graduate student, who would present results, and then get feedback as to what to do for the following week.

Major Topics Discussed / Explored by the Phylogenetic Tree Sub-Group:

- John Aston's phylogenies with covariance matrix representations of dialects as leaves.
- Metrics for covariance matrices as data.
- Susan Holmes gave excellent overview of phylogenetic methods and ideas.
- Phylogenetic Bootstrapping and Multi-Dimensional Scaling
- Discussed Tom Nye's approach to PCA in phylogenetic tree space
- Explored MDS of Brain Vessel trees, and used to probe curvature of tree space. Motivated idea of MDS with
- "curved target space".
- Stickiness fed into heuristic improvement over Sturm's algorithm.
- Visualization of geodesic paths.
- Smoothing in Phylogenetic Tree Space.
- Gene Trees vs. Species Trees.
- Scott Schmidler's phylogenies with shape objects as leaves.
- Variations on Sturm's Algorithm.
- Challenges in defining "convex set".

Major Topics Discussed / Explored by the Dyck Path Sub-Group:

- Correspondence issues, motivated by standard shape considerations.

- Branching processes as probability models.
- Graphical representations of Dyck Paths
- Dyck path mean, and unsuitability of it (some potential fixes)
- Node length representation and analysis
- Tree Pruning Analysis, for deeper insights into population structure.
- Explored transformed Node Length Representations
- Problem with Dyck Path PCA leaving Tree Space. Solution: Non-Negative Matrix Factorization
- Neuromorph Group: Neurons as tree structured data objects.
- Kim's Bayesian Gaussian Latent Factor Model Approach

7.2 Participants

Participants, Phylogenetic Tree Sub-Group: The pivotal graduate student, who did most of the work in this group was Sean Skewerer, UNC Statistics and Operations Research. Active Regular Participants were: J. S. Marron (UNC), Ezra Miller (Duke), Megan Owen (UC Berkeley), John Aston (Warwick), Huiling Le (Nottingham) , Snehalata Huzurbazar (Wyoming). Irregular participants included: Ian Dryden (South Carolina), Susan Holmes (Stanford), Scott Schmidler (Duke), Chris Challis (Duke), Sridevi Polavaram (NeuroMorph).
Participants, Dyck Path Sub-Group: The central graduate student, who did most of the work in this group was Dan Shen, UNC Statistics and Operations Research. Active Regular Participants were: J. S. Marron (UNC), Lingsong Zhang (Purdue), Haipeng Shen (UNC), Yolanda Munoz (Michigan Tech.), Yongdai Kim (Seoul National). Irregular participants included: Haonan Wang (Colorado State) Sarabdeep Singh (Wyoming), Elizabeth Bullitt (UNC), Andrew Wood (Nottingham) , Ruchi Parekh (NeuroMorph).