Submittee: Ryan Budney
Date Submitted: 2015-09-30 13:30
Title: Applied Topology and High-Dimensional Data Analysis meeting in Victoria, 2015
Event Type: Conference-Workshop

Location:
University of Victoria

Dates:
August 17th--28th.

Topic:
This conference aims to bring together applied topologists and statisticians that are curious about each-other's work. We anticipate having two relatively-distinct groups: 1) topologists that have been working with applied algebraic-topological tools like persistent homology, 2) statisticians working in the area of high-dimensional data analysis, and researchers from fields that involve large data problems. We aim to have as many general-interest talks as possible but as talks get more technical and special-interest we will have some parallel sessions aimed towards a particular audience. The other principal aspect of the conference is the first week has talks aimed primarily at a graduate student audience, covering the most basic topics in algebraic topology, persistent homology, and the basics of the fields people are applying these topics to, such as economics, brain imaging and genomics.

Methodology:
Introductory and summary lectures, as well as research lectures, poster sessions, a problem/discussion session.

Objectives Achieved:
One of the main goals of the workshop was to bring together statisticians with applied topologists to allow statisticians the time to get acquainted with persistent homology, what is known and unknown about the subject. This allowed two groups with fairly different world views to get a sense for what each other knows, and what their concerns are. We appear to have a few meaningful collaborations coming out of the workshop (Nathoo and Levanger, for example).

The discussion session was particularly fruitful, if admittedly basic. Much of the conference involved two groups of people stretching to understand each other's work. The discussion session perhaps resolved some of the most basic questions statisticians had about Persistent Homology, and it also brought into focus for topologists what statisticians are curious about with the field of Persistent Homology. To many of the statisticians at the conference, Persistent Homology appears to be a good tool to start resolving foundational issues surrounding assumptions of sparsity of data. Much high-dimensional data analysis assumes sparsity in the underlying data. Persistent homology is a tool which many hope will eventually be a useful tool for potentially verifying or denying sparsity of data.

Scientific Highlights:
It is too early to cite papers (the meeting concluded less than a month ago). But the Levanger-Nathoo collaboration appears to be producing some nice results, merging the ideas of persistent homology and the lasso method. Budney is starting to collaborate with Lesnick, Bubenik and Wright. In my opinion the highlight of the conference was the build-up of the theory of multi-persistence, culminating in Wright's presentation (world premiere) of the Lesnick-Wright software for visualizing multi-variable persistence.

Organizers:
Budney, Ryan, Mathematics and Statistics, University of Victoria
Nathoo, Farouk, Mathematics and Statistics, University of Victoria
Ahmed, Ejaz, Mathematics and Statistics, Brock University.

Speakers:
Ejaz Ahmed (Brock). Title: High Dimensional Data Analysis: Making Sense or Folly. Abstract: In high-dimensional statistics settings where number of variables is greater than observations, or when number of variables are increasing with the sample size, many penalized regularization strategies were studied for simultaneous variable selection and post-estimation. Penalty estimation strategy yields good results when the model at is assumed to be sparse. However, in a real scenario a model may include both sparse signals and weak signals. In this setting variable selection methods may not distinguish predictors with weak signals and sparse signals and will treat weak signals as sparse signals. The prediction based on a selected submodel may not be preferable due to selection bias. We suggest a high-dimensional shrinkage estimation strategy to improve the prediction performance of a submodel. Such a high-dimensional shrinkage estimator (HDSE) is constructed by shrinking a full model estimator in the direction of a candidate submodel. We demonstrate that the proposed HDSE performs uniformly better than the full estimator. Interestingly, it improves the prediction performance of the selected submodel. The relative performance of the proposed HDSE strategy is appraised by both simulation studies and the real data analysis.

Robert Ghrist (UPenn). Title: Local-to-Global Data: Applied Algebraic Topology. Abstract: Many contemporary challenges in the sciences concern the inference of global features from local data. This passage from local- to global- data is as subtle as it is fundamental; however, it is not unprecedented. In the mathematical sciences, several types of local-to-global challenges were overcome with new techniques -- from topology, homological algebra, and sheaves. This talk will outline both the vision and the first steps of exporting homological and topological tools to the data sciences, with an abundance of examples.

Tim Johnson (Michigan). Title: An Overview of the Statistical Analysis of Neuorimaiging Data. Abstract: In this talk I will present an overview of the statistical analysis of neuroimaging data. In particular, I will focus on the analysis of fMRI data including brain mapping and connectivity studies. I will discuss both the merits and pitfalls of the most commonly used methods and models. I will also briefly discuss Random Field Theory (RFT) that is used to address the massive multiple comparisons problem in fMRI analysis. It turns out that the Euler characteristic -- a topological measure -- plays a central role in RFT.

Mat Taddy (Chicago). Title: Big Data and Bayesian Nonparametrics. Abstract: Big Data is often characterized by large sample sizes and strange variable distributions. For example, consumer spending on an e-commerce website will have 10-100s million observations weekly with density spikes at zero and elsewhere and very fat right tails. Such spending will also be accompanied by a

large set of potential covariates. These properties -- big and strange -- beg for nonparametric analysis. We revisit a flavor of distribution-free Bayesian nonparametrics that approximates the data generating process (DGP) with a multinomial sampling model. This model then serves as the basis for analysis of statistics -- functionals of the DGP -- that are useful for decision making regardless of the true DGP. For example, we'll discuss analysis of a least-squares indexing of treatment effect heterogeneity onto user characteristics, as well as analysis of decision trees developed for fraud prediction. The result is a framework for scalable nonparametric Bayesian decision making on massive data.

Ryan Budney (Victoria). Title: Abelian Groups and Smith Normal Form. Abstract: Abelian groups are the language one uses to describe homology. This talk will be about what abelian groups are, how they are related to each other, classified, how one can compute with them and mechanisms for coping with large computations.

Mike Mandell (Indiana). Title: Introduction to Simplicial Homology. Abstract: This talk will introduce the basic definitions and properties of the homology of simplicial complexes.

Joel Hass (UC Davis). Title: Filtered spaces and barcodes. Abstract: Understanding the structure of data at different scales leads naturally to the notion of a filtered spaces. Persistent homology allows us to understand the structure of such spaces in a way that separates noise from real information. The output of a persistent homology computation is a "barcode" that captures when holes of a given dimension are created and when they are filled. We will introduce the concepts of a filtered space and a barcode and give some simple examples.

Michael Lesnick (Columbia). Title: Stability of persistent homology. Abstract: This talk will introduce the stability theorem for persistent homology, along with some extensions. The stability theorem is arguably the central result in the persistence theory: It provides the core mathematical justification for the use of persistent homology in the study of noisy data, and serves as a starting point for the development of statistical foundations for persistence.

Julie Zhou (Victoria). Title: Robust statistics. Abstract: Since outliers may have huge influence on some statistical procedures, it is very important to use robust statistics to do data analysis and detect outliers. In the talk, I will give a short introduction to robust statistics. Two important measures, influence function and breakdown point, will be defined to assess the robustness of statistics. Robust location and scale estimators, robust regression, confidence region, and robust covariance matrix will be discussed. Various examples and R functions for robust procedures will be given.

Vidit Nanda (UPenn). Title: Discrete Morse theory for computing homology. Abstract: From a theoretical complexity perspective, the algebraic question of how one computes homology (persistent or otherwise) of a finite cell complex has a very satisfying solution. However, as topological methods pervade data analysis, one requires significantly faster computation: there is no solace in cubical complexity algorithms when staring down billions of simplices. In this talk, we will examine a discrete version of Morse theory which may be used to whittle a gigantic cell complex down to tractable size without losing any (persistent) homological information.

Rachel Levanger (Rutgers). Title: Using Persistent Homology to study dynamics in the space of persistence diagrams, Part I. Abstract: It is common practice to study dynamical systems on domains with periodic boundary conditions to remove boundary effects imposed by a domain of finite size. While this solves one problem, it potentially creates another: two solutions that are symmetry-related may be seen as separate solutions to the system. While there are classical methods to determine if two solutions are symmetry-related, we show how persistent homology is a natural tool that can be used to quotient out these symmetries. Meant to be an introduction to

persistent homology and applied topology, we assume no background in fluid dynamics, and only minimal familiarity with dynamical systems. Rather, this talk is meant to familiarize the audience with two flavors of applications of persistent homology -- reducing a scalar field to a persistence diagram and analyzing the shape of a point cloud -- and to see how these two methods work together to say something about the dynamics of a time-evolving system.

Justin Curry (UPenn). Title: Clustering with Cosheaves. Abstract: We will begin by recalling how the Reeb graph tracks clusters for a real-valued map. From here we will understand the analogous construction for maps to more general parameter spaces, such as R^n. This will motivate the introduction of stratification theory and the theory of constructible cosheaves, which are equivalent to functors from MacPherson's entrance path category, which I will describe.

Farouk Nathoo (UVic). Title: High-dimensional statistics for neuro-imaging. Abstract: I will discuss three problems involving the analysis of neuroimaging data where the number of unknown parameters is far greater than the number of observations. In the first case I will discuss the neuroelectromagnetic inverse problem that arises in studies involving electroencephalography (EEG) and magnetoencephalography (MEG). This is an ill-posed inverse problem that involves the recovery of time-varying neural activity at a large number of locations within the brain, from electromagnetic signals recorded at a relatively small number of external locations on or near the scalp. Framing this problem within the context of spatial variable selection for an underdetermined functional linear model, we propose a spatial mixture formulation where the profile of electrical activity within the brain is represented through location-specific spike-and-slab priors based on a spatial logistic specification. The prior specification accommodates spatial clustering in brain activation, while also allowing for the inclusion of auxiliary information derived from alternative imaging modalities, such as functional magnetic resonance imaging (fMRI). We develop a variational Bayes approach for computing estimates of neural source activity, and incorporate a nonparametric bootstrap for interval estimation. In the second case I will discuss statistical analysis for Imaging Genomics. Recent advances in technology for brain imaging and high-throughput genotyping have motivated studies examining the influence of genetic variation on brain structure. In this setting high-dimensional regression for multi-SNP association analysis is challenging as the response variables obtained through brain imaging comprise potentially interlinked endophenotypes, and in addition, there is a desire to incorporate a biological group structure among SNPs based on their belonging genes.Wang et al. (Bioinformatics, 2012) have recently developed an approach for the analysis of imaging genomic studies based on penalized regression with regularization based on a novel group l_{2,1}-norm penalty which encourages sparsity at the gene level. While incorporating a number of useful features, a shortcoming of the proposed approach is that it only furnishes a point estimate and techniques for obtaining valid standard errors or interval estimates are not provided. We solve this problem by developing a corresponding Bayesian formulation based on a three-level hierarchical model that allows for full posterior inference using Gibbs sampling. Selection of tuning parameters for the model is discussed in detail and our proposed methodology is investigated using simulation studies as well as through the analysis of a large dataset collected as part of the Alzheimer's Disease Neuroimaging Initiative. In the third case I will discuss the problem of brain decoding, a problem that involves the determination of a subject's cognitive state or an associated stimulus from functional neuroimaging data measuring brain activity. In this setting the cognitive state is typically characterized by an element of a finite set, and the neuroimaging data comprise voluminous amounts of spatiotemporal data measuring some aspect of the neural signal. The associated statistical problem is one of classification from high-dimensional data. We explore the use of functional principal component analysis, mutual information networks, and persistent homology for examining the data through exploratory analysis and for constructing features characterizing the neural signal for brain decoding. These features are incorporated into a classifier based on symmetric multinomial logistic regression with elastic net regularization. The approaches are illustrated in an application where the task is to infer, from brain activity measured with magnetoencephalography (MEG), the type of video stimulus shown to a subject.

Peter Bubenik (UFlorida). Title: Topological Data Analysis and Machine Learning. Abstract: Topology aggregates local metric data to provide a global summary of the "shape" of the data. Persistent homology provides a summary of how the shape of the data changes with respect to changes of a parameter. For example, if the parameter is scale, then persistent homology provides a multiscale descriptor of the data. I will give an introduction of various ways in which topological data analysis can be used to provide features and kernels useful for statistical analysis and machine learning. The necessary addition to previous work is a function from the set of topological summaries to a feature space. More precisely, one requires a map from the set of persistence diagrams to a Hilbert space. From this feature map, one obtains a kernel. Using such features and kernels, one can combine topological data analysis with techniques from statistics and machine learning. These methods provide a principled approach to dimension reduction and visualization helping scientists and engineers to make sense of their big data. I will show how this approach can be used to differentiate two classes of proteins.

Joel Hass (UC Davis). Title: Diffeomorphisms of surfaces and some applications. Abstract: There is a tremendous amount of geometric data being gathered from MRI, ultrasound, scanners, satellites, cameras etc. Much of this data concerns the geometry of surfaces, such as brain cortices, faces, protein surfaces and bones. I will discuss how methods developed in low-dimensional topology and geometry can be used to measure the resemblance of pairs of surfaces. I will show several applications, including alignment of brain images, protein structure prediction, and automatic construction of evolutionary trees.

Vidit Nanda. Title: 2d persistence and protein compressibility. Abstract: A standard question in contemporary proteomics asks which properties of proteins may be directly inferred from their molecular structure. Using only X-Ray crystallography data (of the type which is cataloged in the Protein Data Bank), I will outline a method which accurately estimates the compressibility of a given protein. The method involves imposing a filtered simplicial structure around the atom centers, computing various algebraic-topological invariants, and some rudimentary statistical techniques. This is joint work with Marcio Gameiro, Yasu Hiraoka, Shunsuke Izumi, Miro Kramar and Konstantin Mischaikow.

Kofi Placid Adragni (U.Maryland Baltimore, Math & Stats). Title: Hierarchical Principal Fitted Components Abstract: Sufficient dimension reduction methods are statistical methods designed to reduce the dimensionality of large datasets without loss of regression information. In obtaining the sufficient reduction, it is often assumed that all observations are independent. This is a helpful assumption for estimating a minimal sufficient dimension reduction subspace via inverse regression. However, observations are often recorded on subjects or clusters. While the observations from subject to subject could be independent, the within-subject observations are likely dependent. Treating the within-subject observations as independent may not be appropriate, and may adversely affect the estimation of the central subspace. We propose a method for estimation of the central subspace when the observations are dependent within cluster. The proposed methodology is built upon principal fitted components, a likelihood-based method for sufficient dimension reduction.

Mirza Faisal Beg. Title: Problems in High Dimensional Computational Brain and Eye Anatomy Abstract: Computational tools to study shape change in human anatomy as a function of disease lead to a small number of points in very high dimensional data. I will be presenting some recent work from our group in applying the toolbox of Computational Anatomy in the setting of brain and retina morphometry, and applications in Alzheimer's disease and Glaucoma. If time permits, I will also talk about network-based structural biomarkers as well as a unified voxel-based morphometry/tensor-based morphometry method for generating structural measures for discriminating between different dementias.

Luke Bornn (Simon Fraser). Title: Efficient Representations of Massive Spatio-Temporal Point Processes. Abstract: In this talk I will show how disjoint and compact bases provide a natural and intuitive approach to modeling spatial and spatio-temporal point processes. I will argue that inducing spatially compact bases allows for more intuitive representation of many real-world point processes while allowing computational tractability. I will demonstrate how such a result can be obtained through non-negative matrix factorization, and subsequently extend this idea to a full generative model.

Ivor Cribben (U.Alberta, business). Title: A new method for estimating spectral clustering change points for multivariate time series. Abstract: Spectral clustering is a computationally feasible and model-free method widely used in the identification of communities in networks. In this work, we introduce a data-driven method that detects change points in the network structure of a multivariate time series, with each component of the time series represented by a node in the network. Spectral clustering allows us to consider high dimensional time series where the number of time series is greater than the number of time points (n less than p). The method allows for estimation of both the time of change in the network structure and the graph between each pair of change points, without prior knowledge of the number or location of the change points. The stationary bootstrap is used to perform inference on the change points. The method is applied to various simulated high dimensional data sets as well as to a resting state functional magnetic resonance imaging (fMRI) data set. The results illustrate the method's ability to observe how the network structure between different brain regions changes over the experimental time course. The method promises to offer a deep insight into the inner workings and dynamics of the brain.

Mike Daniels (U.Texas, Austin, Integrative Biology). Title: Sequential BART for imputation of missing covariates Abstract: To conduct comparative effectiveness research using electronic health records (EHR), many covariates are typically needed to adjust for selection and confounding biases. Unfortunately, it is typical to have missingness in these covariates. Just using cases with complete covariates will result in considerable efficiency losses and likely bias. We explore a flexible Bayesian nonparametric approach to impute the missing covariates which involves factoring the joint distribution of the covariates with missingness into their sequential conditionals and applying Bayesian additive regression trees (BART) to model each conditional. Obtaining the posterior for each conditional can be done simultaneously. We provide details on the computational algorithm and compare to other methods, including MICE.

Guoqing Diao (George Mason U. stats). Title: Conditional Variable Screening in High-Dimensional Binary Classification Abstract: Most existing variable screening methods for high-dimensional data rely on strong parametric modelling assumptions that are often violated in real applications. Recently, Mai and Zou (2013) proposed a Kolmogorov filter for high-dimensional binary classification based on the Kolmogorov-Smirnov statistic. This screening method, however, does not account for the effects of potential confounders. We propose a new nonparametric conditional screening method to assess the conditional contributions of the individual predictors in the presence of known confounders. A bootstrap method is proposed to assess the significance for each predictor. The proposed method retains the features of the Kolmogorov filter and is shown to enjoy the sure screening property under the much weakened model assumptions compared to the parametric conditional screening methods. We illustrate the proposed method through extensive simulation studies and real applications. This is joint work with Jing Qin.

Lee Dicker (Rutgers, stats). Title: Efficient variance estimation for high-dimensional linear models. Abstract: We consider high-dimensional linear models with random predictors, and argue that there is little difference between fixed-effects models and random-effects models, under appropriate conditions. In the fixed-effects model, the regression coefficients "inherit" randomness from the predictors and, consequently, methods and results for random-effects models may be ported to the fixed-effects setting. In particular, following an empirical Bayes strategy, we derive high-dimensional

optimality results for ridge regression and asymptotically efficient estimators for the residual variance (σ^2) in high dimensions, in the fixed-effects linear model. High-dimensional residual variance estimation has recently received increased attention in the statistical literature, with important applications in model selection, regression diagnostics, and applied fields, such as genomics. Our results for estimating the residual variance estimation do not require any sparsity assumptions on the regression parameters (which are common in other approaches) and appear to be the first high-dimensional efficiency results of their kind. This is joint work with Murat A. Erdogdu (Stanford)

Kjell Doksum (Wisconsin, Stats). Title: Perspectives on High Dimensional Data Analysis. Randomize, Mix and Match Abstract: The growth of high dimensional data sets has been followed by a growth of methods for analyzing such data sets. One approach is to mix existing methods. For instance, "sparse PCA" mixes PCA and Lasso by replacing the sum of squares condition in PCA is by a condition on the sum of absolute values of coefficients(approximately). Another approach is to repeat a method on randomly selected subsets of variables and then to average or combine the outcomes. This approach is sometimes referred to as "bagging" or "sketching". "Random forest", which in many frameworks improves substantially on the tree approach of CART, is a prime example. Similarly, we find that "random Lasso" typically has smaller prediction error than Lasso. We examine this phenomena, and point to situations where it leads to smaller risk, and to situations where it has limitations. Note that because this "randomize" approach uses subsets of variables, a method that may not be appropriate when the number of variables exceeds the sample size, can now be used for high dimensional data.

Xiaoli Gao (U. North Carolina, Greensboro, math&stats). Title: Penalized adaptive weighted least square regression Abstract: In high-dimensional settings, penalized least squares approach can lose its efficiency in both estimation and variable selection due to the existence of heteroskedasticity. In this manuscript, we propose a novel approach, penalized adaptive weighted least squares (PAWLS), for simultaneous robust estimation and variable selection. The proposed PAWLS is justified from both Bayesian understanding and robust variable selection points of view. We also establish oracle inequalities for both regression coefficients and heterogeneous parameters. The performance of the proposed estimator is evaluated in both simulation studies and real examples.

Yulia Gel (Waterloo, Statistics). Title: Using Data Depth vs. Depth classifier for detecting communities in networks Abstract: We propose a new nonparametric spectral clustering algorithm for detecting communities in large networks using the Depth vs. Depth (DD) classifier. Under the stochastic block model (SBM) and pre-defined number of clusters, we study the performance of four data depth functions, i.e. simplicial depth, half-space depth, Mahalanobis depth, and projection depth. We investigate the effect of regularization on a classification error and compare the new DD classifier on networks with the regularized clustering algorithm based on the K-means approach. It is a joint work with Yahui Tian.

Jinko Graham (Simon Fraser U, stats). Title: Integrative analysis of genomic and neuro-imaging data Abstract: For exploring large and complex data from genomic and neuro-imaging studies, multivariate projection is a valuable tool. In particular, canonical correlation analysis (CCA) has been useful for integrating genomic and neuroimaging data, and understanding their relationship. However, in imaging-genomics studies, the number of features greatly exceeds the number of individuals, and traditional CCA methods cannot be used. Moreover, high correlations between the features de-stabilize analyses. To avoid these problems and facilitate interpretation, sparse CCA (SCCA) has been proposed using an elastic-net-inspired penalty. SCCA reduces dimensionality and facilitates the identification of relevant features during the integration process. To remove additional unimportant features, further filtration based on the adaptive lasso or BIC may be applied. In this ongoing work, we use SCCA to explore the relationship between genome-wide variation at

single-nucleotide polymorphisms (SNPs) and region-specific rates of decline in brain structure and function, as measured by functional magnetic resonance imaging. We apply these methods to data from the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI1), an imaging-genomics study of Alzheimer's disease and mild cognitive impairment. I will describe the statistical challenges encountered so far, and propose possible approaches to deal with them.

Bei Jiang (Columbia University, Biostats). Title: Modeling Placebo Response using EEG data through a Hierarchical Reduced Rank Model. Abstract: There is growing evidence that individual differences among depression patients on Electrophysiology (EEG), fMRI and other brain imaging measurements may be predictive of potential treatment response. In this talk we discuss approaches to identifying potential placebo responders, i.e., a subgroup who benefits sufficiently from inactive drug treatments, using EEG measurements as a matrix (order-2 tensor) predictor. Given the high dimensionality of the problem, we consider a reduced rank regression model with a data-driven regularization. Our approach will be evaluated through simulations and will be applied to data from a large placebo-controlled clinical trial of major depressive disorders.

Timothy D. Johnson (Michigan, Biostats). Title: Analysis of Point Pattern Imaging Data using Log Gaussian Cox Processes with Spatially Varying Coefficients Abstract: Log Gaussian Cox Processes (LGCP) are used extensively to model point pattern data. In these models, the log intensity function is modeled semiparametrically as a linear combination of spatially varying covariates with scalar coefficients plus a Gaussian process that models the random spatial variation. Almost exclusively, the point pattern data are a single realization from the driving point process. In contrast, our motivating data are lesion locations from a cohort of Multiple Sclerosis patients with patient specific covariates measuring disease severity. Patient specific covariates enter the model as a linear combination with spatially varying coefficients. Our goal is to correlate disease severity with lesion location within the brain. Estimation of the LGCP intensity function is typically performed in the Bayesian framework using the Metropolis adjusted Langevin algorithm (MALA) and, more recently, Riemannian manifold Hamiltonian Monte Carlo (RMHMC). Due to the extremely large size of our problem -- 3D data (64 x 64 x 64) on 240 subjects -- we show that MALA performs poorly in terms of posterior sampling and that RMHMC is computationally intractable. As a compromise between these two extremes, we show that posterior estimation via Hamiltonian Monte Carlo performs exceptionally well in terms of speed of convergence and Markov chain mixing properties.

Abbas, Khalili (McGill, math & stats). Post-Model Selection Inference for Finite Mixture of Regression (FMR) Models Abstract: Finite Mixture of Regression (FMR) models are used in situations where several sub-populations within a population exist but the sub-populations themselves are unknown. Recently developed statistical methods for parameter estimation and variable selection in FMR models involve regularization techniques such as the LASSO and SCAD. However, it is well-known that the variability due to model selection stage causes bias in the resulting statistical inference. In this presentation, we address this issue in FMR and several related mixture models. This is a joint work with Anand N. Vidyashankar from Dept. of Statistics at George Mason University.

Peter Kim (Guelph). Title: Phylogenetic LASSO: Pruning the tree of life Abstract: We model the effects of bacterial composition of the human gut microbiome by applying a hierarchical LASSO in series to penalize the variables corresponding to high level organismal classification, in essence pruning the tree of life by phylum, class, etc. Based on Zhu and Zhou's group L^{1/2} hierarchical LASSO, we employ this in seeking the contributors to recovery in faecal microbiota therapy for recurrent Clostridium difficile infection.

Linglong Kong (U. Alberta, Math & Stats). Title: Tensor Approximation in Functional Linear Quantile Regression Abstract: We consider the estimation in functional linear quantile regression in which the dependent variable is scalar while the covariate is a function, and the conditional quantile for

each fixed quantile index is modeled as a linear functional of the covariate. There are two common approaches for modeling the conditional mean as a linear functional of the covariate. One is to use the functional principal components of the covariates as basis to represent the functional covariate effect. The other one is to extend the partial least square to model the functional effect. The former belongs to unsupervised method and has been generalized to functional linear quantile regression. The later is a supervised method and is superior to the unsupervised PCA method. In this talk, we propose to use partial quantile regression and its tensor approximation to estimate the functional effect in functional linear quantile regression. Asymptotic properties have been studied and show the virtue of our method in large sample. Simulation study is conducted to compare it with existing methods. A real data example in stroke study is analyzed and some interesting findings are discovered. Joint work with Ivan Mizera and Dengdeng Yu.

Michael Lesnick (IMA and Columbia) Title: Interactive Visualization of 2-D Persistent Homology. Abstract: In topological data analysis, we often study data by associating to the data a filtered topological space, whose structure we can then examine using persistent homology. However, in many settings, a single filtered space is not a rich enough invariant to encode the interesting structure of our data. This motivates the study of multidimensional persistence, which associates to the data a topological space simultaneously equipped with two or more filtrations. The homological invariants of these "multifiltered spaces," while much richer than their 1-D counterparts, are also far more complicated. As such, adapting the usual 1-D persistent homology methodology for data analysis to the multi-D setting requires some some new ideas. In this talk, I'll introduce multi-D persistent homology and discuss joint work with Matthew Wright on the development of a tool for the interactive visualization of 2-D persistent homology.

Mary Lesperance (UVic, math&stats). Title: A Bayesian Group-Sparse Regression Model with application to Brain Imaging Genomics Abstract: Advances in technology for brain imaging and genotyping have motivated studies examining the relationships between genetic variation and brain structure. Wang et al. (Bioinformatics, 2012) developed an approach for simultaneous regression parameter estimation and SNP selection based on penalized regression with a group l_{2,1}-norm penalty. The group-norm penalty formulation incorporates the biological group structures among SNPs induced from their genetic arrangement and enforces sparsity at the group level. Wang et al. do not provide standard errors or other inferential methodology for their parameter estimates. In this paper, we propose a corresponding Bayesian model that allows for full posterior inference for the regression parameters using Gibbs sampling. Properties of our method are investigated using simulation studies and the methodology is applied to a large dataset collected as part of the Alzheimer's Disease Neuroimaging Initiative.

Rachel Levanger (Rutgers) Title: Using Persistent Homology to study dynamics in the space of persistence diagrams, Part II Abstract: Building on the concepts from my talk in Part I, we address more nuanced issues that occur when studying the dynamics of fluid flows. This includes addressing sampling rates for periodic orbits and techniques for analyzing large point clouds. We also explore the use of diffusion map projections and show how this approach can help to organize and correlate the data to natural measurements of the system.

Yi Li (U.Michigan, Biostats) Title: Covariance-Insured Screening Methods for Ultrahigh Dimensional Variable Selection Abstract: Effective screening methods are crucial to the analysis of big biomedical data. The popular sure independence screening relies on restricted assumptions such as the partial faithfulness condition, e.g, the partial correlation between outcome and covariates can be inferred from their marginal correlation. However, such a restrictive assumption is often violated, as the marginal effects of predictors may be quite different from their joint effects, especially when the covariates are correlated. We propose a covariance-insured screening (CIS) framework that utilizes the dependence among covariates and identify important features that are likely to be missed by marginal screening procedures such as sure independence screening. The proposed framework

encompasses linear regression models, generalized linear regression models, survival models, and classification of multi-level outcomes.

Xuewen Lu (U.Calgary, math & stats) Title: Variable Selection in Log-linear Birnbaum-Saunders Regression Model for High-dimensional Survival Data via the Elastic-Net and stochastic EM Abstract: Birnbaum-Saunders (BS) distribution is broadly used to model failure times in reliability and survival analysis. In this article, we propose a simultaneous parameter estimation and variable selection procedure in a log-linear BS regression model for high-dimensional survival data. To deal with censored survival data, we iteratively run a combination of the stochastic EM algorithm (SEM) and variable selection procedure to generate pseudo-complete data and select variables until convergence. Treating pseudo-complete data as uncensored data via SEM makes it possible to incorporate iterative penalized least squares and simplify computation. We demonstrate the e!cacy of our method using simulated and real data sets.

Shuangge (Steven) Ma (Yale Biostats) Title: Promoting Similarity of Sparsity Structures in Integrative Analysis Abstract: For data with high-dimensional covariates but small to moderate sample sizes, the analysis of single datasets often generates unsatisfactory results. The integrative analysis of multiple independent datasets provides an effective way of pooling information and outperforms single-dataset analysis and some alternative multi-datasets approaches including meta-analysis. Under certain scenarios, multiple datasets are expected to share common important covariates, that is, their models have similarity in sparsity structures. However, the existing methods do not have a mechanism to promote the similarity of sparsity structures in integrative analysis. In this study, we consider penalized variable selection and estimation in integrative analysis. We develop a penalization based approach, which is the first to explicitly promote the similarity of sparsity structures. Computationally it is realized using a coordinate descent algorithm. Theoretically it has the much desired consistency properties. In simulation, it significantly outperforms the competing alternative when the models in multiple datasets share common important covariates. It has better or similar performance as the alternative when there is no shared important covariate. Thus it provides a "safe" choice for data analysis. Applying the proposed method to three lung cancer datasets with gene expression measurements leads to models with significantly more similar sparsity structures and better prediction performance.

Debashis Mondal (Oregon State, stats). Title: Matrix-free computations for Gaussian Markov random fields and related spatial processes Abstract: Since their introduction in statistics through the seminal works of Julian Besag, Gaussian Markov random fields have become central to spatial statistics, with applications in agriculture, epidemiology, geology, image analysis and other areas of environmental science. Specified by a set of conditional distributions, these Markov random fields provide a very rich and flexible class of spatial processes, and their adaptability to fast statistical calculations, including those based on Markov chain Monte Carlo computations, makes them very attractive to statisticians. In recent years, new perspectives have emerged in connecting Gaussian Markov random fields with geostatistical models, and in advancing vast statistical computations. In this talk, I will briefly discuss the scaling limit of lattice-based Gaussian Markov random fields, namely, the de Wijs process that originates in the famous work of George Matheron on gold mines in South Africa. I will then explore how this continuum limit connection holds out further possibilities to fit a wide range of new continuum models by using Gaussian Markov random fields. The main focus of the talk will be on various novel matrix-free computations for these models. In particular, for spatial mixed linear models, I will present novel frequentist residual maximum likelihood inference via matrix-free h-likelihood computations. I will draw applications both from areal-unit and point-referenced spatial data. Part of this talk is based on joint work with PhD student Somak Dutta.

George Michailidis (U.Michigan, stats). Title: Estimation of Multi-Granger Network Causal Models Abstract: Network Granger causality focuses on estimating Granger causal effects from p times series and it can be operationalized through Vector Autoregressive Models (VAR). The latter

represent a popular class of time series models that has been widely used in applied econometrics and finance and more recently in biomedical applications. In this work, we discuss joint estimation and model selection issues of multiple Granger causal networks. The modeling framework captures two different settings: in the first, different types of measurements (e.g. returns and volatility) are measured over the same set of entities (e.g. financial assets), while in the second the same type of measurement (e.g. gene expression) is measured across the same set of entities (e.g. genes), but over different but related diseases. We introduce appropriate structural penalties that capture the multi-Granger structure, discuss estimation issues and illustrate the results on synthetic and real financial data.

Bin Nan (U.Michigan, Biostats). Title: Large covariance/correlation matrix estimation for temporal data. Abstract: We consider the estimation of high-dimensional covariance and correlation matrices under slow-decaying temporal dependence. For generalized thresholding estimators, convergence rates are obtained and properties of sparsistency and sign-consistency are established. The impact of temporal dependence on convergence rates is also investigated. An intuitive cross-validation method is proposed for the thresholding parameter selection, which shows good performance in simulations. Convergence rates are also obtained for banding method if the covaraince or correlation matrix is bandable. The considered temporal dependence has longer memory than those in the current literature and has particular implications in analyzing resting-state fMRI data for brain connectivity studies.

Yingli Qin (U.Waterloo, stats). Title: Testing the order of a population spectral distribution for high-dimensional data Abstract: Large covariance matrices play a fundamental role in high-dimensional statistics. Investigating the behavior of their eigenvalues can reveal informative structures of large covariance matrices. In this paper, we propose to test the number of distinct population eigenvalues, i.e. the order of the Population Spectral Distribution (PSD). The proposed statistic is based upon a series of bias-reduced estimators of PSD moments. We develop the limiting distributions of our test statistic and the moment estimators. We also prove the (n; p) strong consistency of these estimators, which are clearly demonstrated in our simulation study.

Annie Qu (UIUC, stats). Title: Weak Signal Identification and Inference in Penalized Model Selection Abstract: Weak signal identification and inference are very important in the area of penalized model selection, yet they are under-developed and not well-studied. Existing inference procedures for penalized estimators are mainly focused on strong signals. In this paper, we propose an identification procedure for weak signals in finite samples, and provide a transition phase in-between noise and strong signal strengths. We also introduce a new two-step inferential method to construct better confidence intervals for the identified weak signals. Both theory and numerical studies indicate that the proposed method leads to better confidence coverage for weak signals, compared with those using asymptotic inference. In addition, the proposed method outperforms the perturbation and bootstrap resampling approaches. We illustrate our method for HIV antiretroviral drug susceptibility data to identify genetic mutations associated with HIV drug resistance. This is joint work with Peibei Shi.

Speaker: L. Leticia Ramirez Ramirez (ITAM and CIMAT, Mexico). Title: Nonparametric inference on random networks Abstract: We propose a new nonparametric "patchwork" resampling approach on network inference. This procedure is based on the adaptation of "blocking" argument, developed for bootstrapping of time series and re-tiling for spatial data, to random networks. We discuss how this procedure can be used to quantify estimation uncertainty for network statistics that are functions of degree distribution. We present the developed computationally efficient and data-driven cross-validation algorithm for selecting an optimal "patch" size. We illustrate its implementation for inference on simulated random networks, flight and Wikipedia networks.

Matias Salibian-Barrera (UBC, stats). Title: Outlier Detection for Functional Data Using Principal

Components Abstract: Principal components analysis is a widely used technique that provides an optimal lower-dimensional approximation to multivariate observations in mean square error retaining as much information as possible. In the functional case, a new characterization of elliptical distributions on separable Hilbert spaces allows us to obtain an equivalent stochastic optimality property for the principal component subspaces of random elements on separable Hilbert spaces. This property holds even when second moments do not exist. Furthermore, these lower-dimensional approximations can be very useful in identifying potential outliers among high-dimensional or functional observations. In this talk, we discuss the problem of estimating these finite-dimensional approximating linear subspaces robustly. The new class of robust estimators for principal directions is consistent for elliptical random vectors, and Fisher-consistent for elliptically distributed random elements on arbitrary Hilbert spaces. We illustrate our method on two real functional data sets, where the robust estimator is able to discover atypical observations in the data that would have been missed otherwise. Through a simulation study, we also study its performance when used to detect outlying observations.

Ali Shojaie (U. Washington, Seattle, Biostats). Title: A Significance Test for Graph-Constraint Estimation Abstract: External information on associations between covariates could be helpful in making more accurate statistical estimations. Such information is usually presented using graphs. Those graph-guided estimation methods may also be able to perform automatic variable selection. However, there is no available inference method that calculates p-values for graph-guided estimates. In this paper, we present a testing framework called the Grace test. Grace test could produce coefficient estimates and corresponding p-values with incorporated external graphical information. We show in theory that our method asymptotically controls the type-I error rate regardless of the choice of the graph. However, the power of the test would benefit from using an informative graph. We further propose a more general Grace-ridge test that delivers a higher power than the Grace test when the choice of the graph is not informative. We show via simulations that with a reasonably informative graph, we achieve a substantial gain in power, compared to several existing inference methods that ignore the external graphical information.

Peter X.K. Song (U.Michigan, Biostats). Title: FLAPO: Fused Lasso with the Adaptation of Parameter Ordering -- An approach to parameter fusion in merging multiple studies with repeated measurements Abstract: Combining multiple studies is frequently undertaken in biomedical research to increase sample sizes for statistical power improvement. We consider the marginal model for the regression analysis of repeated measurements collected in several similar studies with potentially different variances and correlation structures. It is of great importance to examine whether there exist common parameters across study-specific marginal models so that simpler models, sensible interpretations and meaningful efficiency gain can be obtained. Merging multiple studies using the classical means of hypothesis testing involves a large number of simultaneous tests for all possible subsets of common regression parameters, in which it results in unduly large degrees of freedom and low statistical power to assess covariate effects. We develop a new fused lasso method, using estimated parameter ordering, to scrutinize only adjacent-pair parameter differences, leading to a substantial reduction for the number of needed comparisons. Our method enjoys the oracle property as does the full fused lasso that involves all pairwise parameter differences. We show that the proposed procedure has smaller error bounds and better finite sample performance than the full fused lasso. We illustrate our method through extensive simulation studies as well as an analysis of merging HIV surveillance cohorts collected over five geographic regions in China, in which the presence or absence of common covariate effects are reflective to relative effectiveness of regional policies on HIV control and prevention. Work joint with Fei Wang and Lu Wang.

Matt Taddy (Chicago, business) Title: Bayesian and empirical Bayesian forests Abstract: We derive ensembles of decision trees through a nonparametric Bayesian model, allowing us to view random forests as samples from a posterior distribution. This insight motivates a class of Bayesian

forest (BF) algorithms that yield small gains in performance and large gains in interpretability. Based on the BF framework, we are able to show that high-level tree hierarchy is stable in large samples. This leads to an empirical Bayesian forest (EBF) algorithm for building approximate BFs on distributed massive datasets and we show that EBFs outperform sub-sampling based alternatives by a large margin.

Anand N.Vidyashankar (George Mason U, Statistics) Title: Big Data in Healthcare: Privacy vs Statistical Efficiency Abstract: Over the past few decades, healthcare has experienced a proliferation of databases associated with a massive collection of medical information. Combinations of such information tend to lead to privacy and security risk. On the other hand, statistical analyses of such combined information is necessary to develop products and services for patient benefits. Thus a natural question is: is it possible to provide both privacy and statistical guarantees when working with big data. In this presentation, we provide new metrics to measure privacy and security risk and develop data analyses techniques accounting for privacy risk. In the process, we illustrate a trade-off between statistical efficiency and privacy. Using this, we develop an algorithm for decision making within a healthcare organization. Our results showcase how healthcare organizations can leverage modern statistical techniques to simultaneously address the twin challenges of creating business value and addressing regulatory guidelines that are emanating from laws like HIPAA/HITECH.

Sijian Wang (U.Wisconson, Madison, Biostats) Title: Integrative analysis of high-dimensional genomic data Abstract: Two types of integrations of multiple genomic datasets are considered in this talk. The first type of integration is based on the datasets from multiple studies. For example, an increasing amount of gene expression datasets for similar biomedical problems is available through public repositories. Integrating data from different but independent studies may facilitate discovery of new biological insights. We propose a meta-lasso method for gene selection with multiple expression data. Through a hierarchical decomposition on regression coefficients, our method not only borrows strength across multiple datasets to boost the power to identify important genes, but also keeps the selection flexibility among datasets to take into account data heterogeneity. The second type of integration is based on the datasets of multiple types. Recently, many genome-wide datasets capturing somatic mutation patterns, DNA copy number alterations, DNA methylation changes and gene expression are simultaneously obtained in the same biological samples. These samples render an integrated data resolution that may not be available with any single data type. We propose an iCluster+ method for pattern discovery (clustering) by integrating diverse data types. The core idea is motivated by the hypothesis that diverse molecular phenotypes can be predicted by a set of orthogonal latent variables that represent distinct molecular drivers, and thus can reveal tumor subgroups of biological and clinical importance.

Matthew Wright (IMA, Minnesota) Title: Euler Characteristic and Data Analysis Abstract: Euler characteristic is a simple topological invariant that is useful for data analysis. In particular, Euler characteristic extends to a topological integration theory for real-valued functions. I will give an introduction to this theory of Euler integration and demonstrate intriguing applications to topological enumeration problems. The observation that Euler characteristic is one of the intrinsic volumes (which appear in the classic Hadwiger Theorem) prompts generalizations of this integration theory, which I will mention. I will also describe recent work applying Euler characteristic to data analysis, especially in the areas of sensor networks, image processing, and object recognition.

Matthew Wright (IMA, Minnesota). Title: Computing Multidimensional Persistent Homology Abstract: Multidimensional persistent homology is highly relevant in various applications in which data is simultaneously filtered by two or more parameters. However, the algebraic complexity of multidimensional persistence modules makes it difficult to extract useful invariants in this setting. In this talk, I will describe recent work with Mike Lesnick to efficiently compute and visualize both the multigraded Betti numbers and the rank invariant of multidimensional persistence modules. In

addition, I will demonstrate a new software program for computing and visualizing these invariants. This talk will follow Mike Lesnick's talk from Tuesday, focusing on computational, rather than theoretical, aspects of our work.

Ruben H. Zamar (UBC, Statisitcs). Title: Robust and sparse hierarchical clustering Abstract: We consider the problem of hierarchical clustering when the dataset has a large number of variables and a relatively small number of observations. Non-informative noise variables and outlying observations are likely to occur in this context. The inclusion of non-informative noise variables may impede the finding of the underlying clusters. Sparse hierarchical clustering (SHC), which uses a subset of adaptively chosen informative variables, outperforms classical methods based on all the variables. However, in the presence of outliers, both classical methods and SHC yield poor results. We propose two approaches for robust sparse hierarchical clustering (RSHC) to deal with both noise variables and outliers. The proposed RSHC methods are based on Tau-estimator and Lasso-type penalty. The first approach is a direct robustification of SHC. The second approach introduces a fast robust version that scales better with the number of observations. Experiments with both simulated and real datasets show that the RSHC approaches work well on clean data and outperform the non-robust alternatives when the dataset contains outliers.

Links:
http://rybu.org/appliedtopmeeting
