Research Statement
James Sharpnack

The prevalence of massive datasets is transforming the field of statistics as we attempt to simultaneously address the statistical efficiency and computational complexity of statistical methods. The modern explosion of data, viz. internet data, sensor networks, genomic microarrays, and financial data, requires that statistical methodology adapts to the assumptions and structure unique to these problems, while ensuring statistical and computational soundness. Much of the recent developments in high-dimensional statistical theory addresses the estimation of mean parameters under sparsity assumptions. My research departs from this framework by considering complex heterogeneous structural assumptions, that appear in modern statistical applications such as sensor networks. Throughout my work, I focus on the development of algorithms with provably low statistical risk and low computational complexity to new statistical problems with complex structural and modeling assumptions.

Structured Tests for Normal Means

High-dimensional signal detection refers to the fundamental statistical task of determining if a given vector of measurements is merely noise or if there is some signal amidst the noise. This is the theoretical framework for detection in low-powered sensor networks, where abnormal measurements indicate the presence of a phenomenon that is being monitored. Other applications are disease outbreak detection, biomedical imaging, differential gene expression, environmental monitoring, and malware detection over a computer network. One common modelling assumption is that the coordinates of the observed vector are independent, normally distributed, with known variance, and under the null hypothesis the mean of every coordinate is zero and under the alternative hypothesis it is non-zero. When a structural assumption is made about the alternative hypothesis, such as the mean vector lies within a known subspace, union of subspaces, or proper subset, of the ambient vector space, one can use statistical tests that target signals of this type which increases the statistical power of the test. This problem has been comprehensively studied when the structural assumption is that the signal lies within Besov and Sobolev spaces [Spokoiny et al., 1996, Ingster and Suslina, 2003], for sparse and weakly supported signals [Donoho and Jin, 2004], and rectangles when coordinates represents points in a grid [Arias-Castro et al., 2005]. The motivating application for the latter is when sensors are deployed in a grid-like fashion, as is the case for pixels in a CCD or CMOS camera, and we are trying to detect a region of abnormally large sensor measurements.

A natural statistic for detecting anomalous sensor readings compares the normalized sum (Z-score) of the sensor measurements within all rectangular regions in the sensor grid to a threshold, which is adjusted with a multiple hypothesis testing correction. In [Sharpnack and Arias-Castro, 2014], we construct a test that adapts this threshold to the size of the rectangle. We characterize the asymptotic distribution under the null and alternative for the adaptive test and the standard scan statistic (with a constant threshold) using recent results on the boundary crossing probability of locally stationary Gaussian random fields. We further constructed an algorithm that achieves these same asymptotic distributions, but with a near-linear (linear up to logarithmic factors) computation time, and it has a simple distributed implementation. Distributed frameworks generally allow for significant reductions in computation time but at the expense of the communication complexity, in the sense that the number of bits passed between relays may be prohibitively large.
The next stage for this work is to understand the effect that communication constraints have on the statistical performance of the scan statistic, by studying this problem from an information theoretic perspective.

**Detecting Graph Structured Normal Means**

The aforementioned problem addresses detecting the mean of a normal vector when the coordinates constitute a grid of regularly spaced points. In sensor network applications, the sensors are not generally arranged in a grid, but form a graph, where edges in the graph can indicate any sort of relationship (like proximity or point-to-point communication). Encoding the structural assumptions for the alternative distribution via a graph has several advantages. Because we seek to make statistical guarantees for any graph structure, the results are not necessarily dependent on the graph taking a specific form, such as being a draw from a specific random graph model. In mobile sensor networks, locating a sensor can be a difficult task, while our algorithms do not require sensor locations but only the edges and their weights which can often be determined by point-to-point sensor communication. For many applications, such as internet data and gene-network analyses, networks can be extracted from various data sources that can be used to structure the detection problem.

In my Ph.D. thesis, I studied a composite alternative hypothesis which specifies that there is an unknown active cluster of vertices (vertices with a non-zero signal size) that is easy to cut out of the graph, in the sense that the edges that connect the cluster to its complement are few. Due to the combinatorial nature of the alternative hypothesis, the scan statistic, which tests individually each such cluster is computationally infeasible. In [Sharpnack et al., 2013b], we provide a computationally feasible relaxation called the spectral scan statistic of the combinatorial optimization and characterize its statistical performance using spectral graph theory. In [Sharpnack et al., 2014], we have modified the spectral scan statistic to construct the graph Fourier scan statistic, which employs a linear filter that performs shrinkage in the graph Fourier domain defined by eigenvectors of the Laplacian operator. We have also proposed and characterized a test statistic that adapts to the unknown cut size parameter. This test enjoys sharper theoretical guarantees than the spectral scan statistic and is nearly optimal for certain graph structures, such as balanced trees. These works demonstrate the connection between spectral graph theory and signal detection over graphs.

While spectral relaxations are nearly-optimal for tree-like graphs, they fail to achieve optimal performance on important graphs such as the grid graph and some geometric random graphs. In [Sharpnack et al., 2013c], we consider a sharper relaxation of the scan statistic based on submodular optimization, called the Lovász extended scan statistic (LESS), and characterize its statistical performance using electrical network theory. We characterized the statistical difficulty of this problem with information theoretic lower bounds and demonstrate that it is nearly-optimal for many geometric random graphs. While the LESS can be implemented using graph cuts, I have developed a path algorithm in [Sharpnack, 2013] that solves the program for all tuning parameter values. These detection algorithms are summarized in [Sharpnack and Singh, 2013]. Collaborators at the Ohio State University (Michael Sharpnack and Kun Huang) and I are currently in the process of using the LESS to perform tests for differential expression in RNA-seq data using protein-protein interaction networks.
Signal Processing over Graphs

We will depart from normal means testing and speak more broadly about signal processing over graphs. The estimation (or reconstruction) of signals over graphs is of fundamental importance to many applications including spatial data with irregular domains (see [Shuman et al., 2013] for an introduction). To this end, we developed a projection estimator that has low mean square error when the signal is smooth over the graph [Sharpnack and Singh, 2010]. Because of the theoretical limitations of linear filters, we decided to develop a wavelet basis, based on spanning trees, over graphs that has approximation theoretic guarantees with a computationally efficient construction. We have shown that if one constructs the spanning tree using random walks in order to form the wavelet basis, it can be used to detect and reconstruct regions of anomalous activity within graphs [Sharpnack et al., 2013a]. With our wavelet basis, we have developed a compressed sensing technique over graphs with active measurements [Krishnamurthy et al., 2013]. This algorithm has been shown to have stronger statistical guarantees than the previous passive counterpart. I intend to expand on this work by developing other wavelet bases designed to approximate specific function classes over graphs. Because multiresolution wavelet transforms have distributed implementations, this work could be used for nearly real-time signal localization schemes in sensor networks.

Total variation denoising, also known as the fused lasso, is a popular method in computer vision, and algorithmically it has been well studied (see [Hoefling, 2010]). The idea is that like the lasso, if we minimize a sum of squares objective with a sparsity inducing penalty, then we can hope to perform correct model selection. In this setting, the sparsity inducing penalty is based on a discrete version of the derivative over a graph. The effect is that signals that are reconstructed are piecewise constant over the graph. We have used spectral graph theory to analyze support recovery for the total variation denoising in a Gaussian model [Sharpnack et al., 2012]. I am currently working on providing mean-square error guarantees for total variation denoising using some recently developed methodology.

Trend filtering is a generalization of the fused lasso, that uses higher-order derivatives to create the sparsity inducing penalty (see [Tibshirani et al., 2014]). In [Wang et al., 2014], we define trend filtering over graphs and study mean-square error recovery and provided several applications to spatial datasets. Because of the generality of graph trend filtering, we are still learning how to characterize its statistical performance for every order of the derivative operator. There are many extensions of this work to be done, such as accommodating compressed measurements, extending it to graph structured time series observations, and applying it to various applications in spatial statistics. There is a strong connection between trend filtering and numerical solutions to partial differential equations, where trend filtering can be interpreted as a data-driven locally adapted finite difference method. I would like to explore this connection and hope to make significant contributions to spatial statistics through this line of research.

High-dimensional Model Selection

I am involved in other projects studying model selection in high-dimensional models. Heteroscedastic models have been extensively studied in the classical statistical literature because of the prevalence of non-identical sample variances in real-world data. While some work has been done to provide high-dimensional mean parameter estimates in sparse linear models that can accommodate heteroscedastic errors, little work has been done to perform model selection for the variance parameters. With this in mind, in [Kolar and Sharpnack, 2012], we have developed a method, that we call the heteroscedastic iterative
penalized pseudolikelihood optimizer (HIPPO), for estimating sparse mean and variance parameters in a high-dimensional heteroscedastic model. This work has important applications in computational finance and macroeconomics, since with HIPPO one can help explain the causes of volatility in markets. We have recently submitted an updated version, [Sharpnack and Kolar, 2014], that greatly improves the theoretical properties known by modifying the HIPPO algorithm. In these works, we have discovered that by taking a pseudolikelihood approach, we can provide variance parameter estimates that perform asymptotically as well as if we had known the true sample means. The HIPPO algorithm uses non-convex penalties and the statistical guarantees are for a local-minimizer (as opposed to global minimizers) of the penalized likelihood. To improve upon this limitation, we are working to extend our analysis to $\ell_1$ penalties. Also, in [Sharpnack and Kolar, 2014], we assumed that the variances have a log-linear functional form, which we intend to generalize this using semi-parametric estimation and generalized additive models.

I am also leading a project with Jelena Bradic to infer graphical model structure from counting processes that are modelled by an infection network. The methodology that we will develop has a promising application to learning functional brain connectivity from neural spike train data. This project allows for very general statistical assumptions and will improve our understanding of model selection in event history data. In the future, I hope to explore graphical model structure learning of counting processes in spatial data and its relation to graph structured inference.

**In Conclusion**

The following is a summary of my future work and ongoing projects.

1. **Information theoretic limits of distributed detection** in sensor and relay networks (with Ery Arias-Castro).
2. **Differential expression in RNA-sequencing**: apply graph scan statistics to testing for differential RNA expression (with Michael Sharpnack and Kun Huang).
3. **Wavelet constructions over graphs and point clouds** with the express goal of providing approximation theoretic and statistical guarantees (with Ery Arias-Castro and Aarti Singh).
4. **Total variation denoising on graphs**: approximate recovery and changepoint localization.
5. **Trend filter over graphs** and other extensions of multivariate trend filtering (with Ryan Tibshirani, Alessandro Rinaldo, Yu-Xiang Wang, and Alex Smola).
6. **Variance function estimation** with $\ell_1$ penalties and other variance functional forms (with Mladen Kolar).
7. **Counting process graph structure** learning from infection networks and event history data (with Jelena Bradic).

I am fortunate to have many great collaborators, including Aarti Singh, Alessandro Rinaldo, Akshay Krishnamurthy, Ery Arias-Castro, Mladen Kolar, Ryan Tibshirani, Yu-Xiang Wang, Alex Smola, and Jelena Bradic. In the future, I intend to build strong collaborations with practitioners in other fields that can use
the statistical methods that I have developed. I have worked with two junior graduate student, Akshay Krishnamurthy and Yu-Xiang Wang, that have resulted in highly productive collaborations and also mentored an undergraduate student in a web application project. These experiences have made me appreciate how mentoring can help my research develop, and have taught me how rewarding and productive it can be to help students pursue their vision for a research project. I am confident that my research and collaborations will continue to be fruitful.

References


6