

VIGRE Fellows 2007-2008

Nicholas Basch

VIGRE Committee Assignment - PIMS-VIGRE (Gunther Uhlmann chairs)

Jennifer Chunn

- (1) Develop Statistics Teaching Assistant Handbook with Will Kleiber

Will Kleiber and I are working on developing a small handbook for statistics Teaching Assistants with various sections including teaching resources available to them (e.g., tutor center, head T.A., CIDR), previous course websites that are available, helpful teaching aids and exercises (e.g., Rice Virtual Lab simulation website to illustrate sampling distribution <http://onlinestatbook.com/rvls.html>).

- (2) Serve on the Professional Development forum committee

The purpose of this committee is to brainstorm, organize, advertise, and put on professional forums for undergraduate and graduate Math/Applied Math/Statistics students who often have questions and need advice on life as a Mathematician/Statistician and the steps to those careers. Topics include: graduate school funding, life as a post-doc, academic career process.

- (3) Learning Initiative project participant

This project began last year with a goal to determine, and make available, learning objectives for the courses offered by the Statistics Department and also evaluate the undergraduate program. My role in the continuation of this project will be to attend and participate in meetings and discussions. Note, my role on this project will be different this year as compared to last spring when I was funded as an RA and this summer when I volunteered my time on this project.

Research Projects and Coursework:

- (1) Probabilistic Mortality Projection Project with Adrian Raftery

Every two years, the United Nations publishes estimates and forecasts of mortality rates for every country in the world. We aim to develop a method to estimate and project mortality rates for every country in the world and assess the uncertainty of those estimates.

- (2) Autumn 2007 quarter -- academic schedule:

Stat 570 - Applied Statistics
BioStat 578 - Spatial Epidemiology
Stat 516 - Stochastic Modeling
Stat 600 - Research with Adrian Raftery
Stat 579 - Data Analysis (auditing)

Krista Gile
August 2008 – PhD
Oxford University
Postdoctoral Prize Research Fellowship

Dissertation Title – “Inference from Partially-Observed Network Data”

As a VIGRE fellow, I am a member of the Undergraduate Seminar Committee. In Autumn 2007, this committee met to begin planning the Undergraduate Seminars, which will begin in Winter 2008. In particular, I have volunteered to be in charge of maintaining and coordinating the schedule of speakers. In this role, I have initiated contact with several presenters for the winter quarter, and put in place a mechanism for coordinating speaker scheduling.

In Winter 2008, my primary VIGRE-related activity was to help organize the Undergraduate Mathematical Sciences Seminar Series. My particular role was to schedule the seminar speakers. We had a variety of speakers from both within and outside the University of Washington, and representing statistics, math, applied math, and several related disciplines.

I also continued my vertical and horizontal integration with the cross-disciplinary UW Network Modeling working group. I also continued my horizontal integration with students of the Center for Statistics and the Social Sciences (CSSS) disciplines through a presentation at the CSSS Student Seminar Series Social. Furthermore, I prepared to extend my vertical integration in the Spring quarter, as I began preparations for teaching the STAT 220 (Basic Statistics) course.

Articles:

Paper: "Modeling Social Networks from Sampled Data"
Authors: Mark S. Handcock and Krista J. Gile
R&R: Annals of Applied Statistics

Abstract:

Network models are widely used to represent relational information among interacting units and the structural implications of these relations. Recently, social network studies have focused a great deal of attention on random graph models of networks whose nodes represent individual social actors and whose edges represent a specified relationship between the actors. Most inference for social network models assumes that the presence or absence of all possible links is observed, that the information is completely reliable, and that there are no measurement (e.g. recording) errors. This is clearly not true in practice, as much network data is collected through sample surveys. In addition even if a census of a population is attempted, individuals and links between individuals are missed (i.e., do not appear in the recorded data). In this paper we develop the conceptual and computational theory for inference based on sampled network information. We first review forms of network sampling designs used in practice. We consider inference from the likelihood framework, and develop a typology of

network data that reflects their treatment within this frame. We then develop inference for social network models based on information from adaptive network mechanisms. We motivate and illustrate these ideas by analyzing the effect of link-tracing sampling designs on a collaboration network.

Paper: "Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models"

Authors: Marijtje A.J. van Duijn, Krista J. Gile, and Mark S. Handcock

R&R: Social Networks

Abstract:

The statistical modeling of social network data is difficult due to the complex dependence structure of the tie variables. Statistical exponential families of distributions provide a flexible way to model such dependence. They enable the statistical characteristics of the network to be encapsulated within an exponential family random graph (ERG) model. For a long time, however, likelihood-based estimation was only feasible for ERG models assuming dyad independence. For more realistic and complex models inference has been based on the pseudo-likelihood. Recent advances in computational methods have made likelihood-based inference practical, and comparison of the different estimators possible.

In this paper, we compare the bias, standard errors, coverage rates and efficiency of maximum likelihood and maximum pseudo-likelihood estimators. We also propose an improved pseudo-likelihood estimation method aimed at reducing bias. The comparison is performed using simulated social network data based on two versions of an empirically realistic network model, the first representing Lazega's law firm data and the second a modified version with increased dependency. We consider estimation of both the natural parameters and the mean-value parameters.

The results clearly show the superiority of the likelihood-based estimators over those based on pseudo-likelihood. The use of the mean value parameterization provides insight into the differences between the estimators and when these differences will matter in practice.

Paper: "Model-based Assessment of the Impact of Missing Data on Inference for Networks"

Authors: Krista J. Gile and Mark S. Handcock

Abstract:

Most inference using social network models assumes that the presence or absence of all relations is known. This is rarely the case. Most social network analysis ignores the problem of missing data by including only actors with complete observations.

In this paper, we use a statistical model for the underlying social network to demonstrate that the computationally parsimonious complete case approach can lead to different conclusions from an approach utilising all observations. We also show that the overall fit to the data is improved by extending the model to represent differences between respondents and non-respondents.

The ideas are motivated and illustrated by an analysis of a friendship network from the National Longitudinal Study of Adolescent Health.

Paper: "Modeling Contact Tracing Data"
Authors: Krista J. Gile and Mark S. Handcock

Abstract:

Contact Tracing is often used as an intervention to find and treat infected people and to head off a potential epidemic. Persons presenting with a reportable disease are asked to name others with whom they have had contact. In the course of this intervention, data are collected on the network structure in the population under study. This paper considers treating these data as a network sample and using them, and other data incidental to the contact tracing process, to make inference about network structure, disease transmissibility, and ultimately, epidemic potential. This work is most applicable to sexually transmitted diseases and sexual networks, where networks are sparse and relations fairly well-defined. We also consider the information benefits of extensions to current contact tracing protocols. We present a joint model for contact and disease structure, and demonstrate computational strategies for fitting the model in the presence of large amounts of unobserved data.

If you'd like you can also include titles for two additional paper that will come out of my dissertation (co-authored with Mark, of course). I don't have abstracts for them yet:

"Evaluating the Assumptions of Respondent-Driven Sampling"

"Improved Estimation for Respondent-Driven Sampling"

William Kleiber

Winter 2008 - This past quarter my VIGRE related activities included a reading course with Tilmann. In particular, the course was initially geared towards spatial statistics, and involved reading various sections of books and current academic journal articles. Gradually, the course evolved into a small research project regarding the construction of cross-correlation functions for multivariate spatial data. Over the next quarter I plan on applying this new method to some data from the Probcast group.

With respect to my VIGRE project, the TA handbook, Jennifer and I have decided to make the handbook available online, and it will include a series of useful hyperlinks. We will meet with Chris Green (or someone closely involved with the department website) to decide the best format for such a site, and will organize important information online, making it easier to access.

Alex Lenkoski

Publications in progress

"Bayesian structural learning and estimation in Gaussian graphical models"

by Alex Lenkoski, Helene Massam and Adrian Dobra

We propose a new stochastic search algorithm for Gaussian graphical models called the mode oriented stochastic search. Our algorithm relies on the existence of a method to accurately and efficiently approximate the marginal likelihood associated with a graphical model when it cannot be computed in closed form. To this end, we develop a new Laplace approximation method to the normalizing constant of a G-Wishart distribution. We show that combining the mode oriented stochastic search and with our marginal likelihood estimation method leads to excellent results with respect to other techniques discussed in the literature. We also describe how to do inference through Bayesian model averaging based on the reduced set of graphical models identified.

"Modeling Growth Determinant Uncertainty Using Gaussian Graphical Models"

by Adrian Dobra, Theo Eicher, and Alex Lenkoski

Bayesian model averaging in linear regression models has become central in the identification of factors that affect cross-country economic growth. In this paper we introduce Gaussian graphical models as a tool for further refining the proposed set of growth determinants. We describe a comprehensive Bayesian framework for performing structural learning in linear regressions, Gaussian graphical models and Gaussian directed acyclic graphs. We discuss jointness measures related to these models. We illustrate our methodology using a dataset involving 41 potential growth factors.

This quarter I was fortunate to be covered by a VIGRE grant. Inside the VIGRE program I was one of four members of the "problem of the week" team. Along with my fellow group members, I was responsible for answering emails regarding problems, logging correct answers and choosing an ice cream winner each week. These responsibilities demanded a small amount of time, especially because of Jonathon Cross's excellent leadership.

The great benefit of the VIGRE program is that it has given me considerable flexibility to pursue work that contributes to the maths and statistics communities at the UW. I feel I have used this time in the spirit of the VIGRE grant. I have continued a major research project with Adrian Dobra on computational methodology, which we are presently wrapping up. Furthermore, we have begun an applied research project on gene pathways that is exciting and of interest to the scientific community at the UW. We have also begun two projects in macroeconomic modeling, the first with Theo Eicher of the economics department and the second with both Theo and Adrian Raftery. These projects are both exciting and extremely relevant to the issues facing economists today.

While the flexibility afforded by a VIGRE grant is beneficial to my research goals and allows me to contribute to the research of the broader UW community, I do feel it is important

as a graduate student to contribute to the instruction of statistics. Luckily, the statistics department has a drop-in tutor center and I have remained a dedicated member of the tutoring staff. Now in my third consecutive quarter at the tutor center, I feel I have gained a good understanding of the majority of undergraduate statistics courses taught at the UW. Such a knowledge-base is indispensable and my available time at the tutor center would be greatly diminished were I not covered by a VIGRE grant.

Larissa Stanberry

VIGRE Committee Assignment - Distinguished Lecturer

August 2008 – PhD
University of Bristol
Lecturer

Dissertation Title – Statistical solutions to some problems in medical imaging

Abstract:

Medical professionals and researchers used a variety of imaging techniques in their clinical practice and scientific investigations. In this talk I will focus on Mammography which is used for breast examinations and routine breast cancer screening. While the mammographic images proved to be a useful non-invasive tool for clinical monitoring, the images often lack detail and clarity. For example, in addition to having limited spatial resolution, skin-air boundary of the imaged breast is often obscured. This boundary is, however, an important initial step in the breast density estimation. Breast density, defined as a proportion of the breast tissue that appears bright on the image, was shown by various research groups to be strongly associated with the risk of breast cancer. In this work we introduce the algorithm to address the boundary detection issue, the first step in density estimation problem. The performance of the method will be demonstrated on the simulated data. We then show the boundary recovery results for the mammogram images and discuss its advantages and possible improvements.

Publications in progress:

L. I. Stanberry and J. Besag. A stochastic method for boundary restoration.

Abstract:

In image analysis, it is often desirable to reconstruct the boundary of an object in a noisy image. For example in mammography, boundary outline of the breast tissue is required as a part of a routine image assessment procedure to determine the risk of developing breast cancer. Various methods for boundary detection have been proposed in the literature, including Markov random field models, active contour models, and stochastic models with polygonal shapes.

This work introduces a novel boundary recovery method based on B-spline curves. The method allows us to estimate a variety of complex contours, including boundaries of non-convex sets

and non-smooth domains. The method is Bayesian and relies on Markov chain Monte Carlo simulations to draw inference about the boundary.

The method is illustrated on the simulated data and applied to recover a skin line on a digitized analog mammogram image. We also discuss the selection of an appropriate loss function.

This is a joint work with Julian Besag. The authors would like to thank Stephen Duffy and Ruth Warren for introducing them to the problem and providing the data.

L.I. Stanberry. The expectation of random sets via oriented distance functions.

Abstract:

In image analysis, it is often desirable to reconstruct the boundary of an object in a noisy image. For example, in mammography, boundary outline of the breast tissue is required as a part of a routine image assessment procedure to determine the risk of developing breast cancer. The boundary can be reconstructed using a Bayesian approach. However, Monte Carlo sample boundaries are structured geometric sets and constructing posterior estimates from the acquired sample curves/surfaces is not straightforward. This paper gives an alternative definition of an expected set as a level set of the expected oriented distance function. The mean set, constructed in such manner, is shown to have a number of desirable properties. In addition to inclusion properties, the expectation is homothetically- and translation invariant. The invariance also holds for the group of orthogonal transformations. The paper introduces a special class of oriented distance functions and establishes for this class a connection between the expectation of a set and the expectation of a random variable from a probability space, on which the random set is defined. The proposed definition and its characteristics are studied in the series of simulated examples and is applied to construct a posterior estimate of the tissue boundary in a mammogram image.