

Statement of Professional Goals and Objectives

Ranjan Maitra

Introduction

The past two decades have seen the growing importance of statistical thought in providing practical solutions to real-world problems. The engineer, the scientist, the medical researcher, the entrepreneur and public policy-maker all want the statistician as a partner in the effort to address vexing issues of practical significance. At play are multiple levels of statistical expertise – that of the experimenter, the analyst or the practitioner. The need for the researcher is then crucial, but no less is that of the educator who nurtures and trains statistically sound scientific lines of inquiry, or of the professional providing valuable service in advancing the goals of the discipline to make research and education possible. In this document, I illustrate my specific goals in each of these areas, my attempts towards realizing them as a member of the faculty in the Department of Statistics at Iowa State University and the statistical community, and my vision for the future.

Research Objectives

Technological and scientific advances have resulted in automated and efficient data collection methods. They have also thrown up interesting research questions in medicine, information, environmental and other sciences. Classical statistical methods have proved to be deficient for ready application in many cases, especially in the context of massive databases. My primary research interests continue to be in the development of practical but statistically sound methodology to address the challenges in such scenarios, with special emphasis on medical image analysis, data mining and the environmental sciences.

My initial exposure to problems of this magnitude was in the context of medical image analysis as a graduate student at the University of Washington. My expertise broadened further with my first position as a Research Scientist in charge of developing the data mining efforts at Bell Communications Research (Bellcore). I continued developing methodology in these areas during my first faculty appointment in the Department of Mathematics and Statistics at the University of Maryland, Baltimore County (UMBC), and, after a two-year slowdown owing to unfortunate double arm fractures, diversified further with my move to a faculty position in the Department of Statistics at Iowa State University. My research now broadly covers the areas of statistical computing and computationally intensive statistical methods for massive datasets, with specific application to clustering and classification, simulation and medical imaging.

A focal point of research is the issue of finding similar groups of records in data which is a very difficult problem that is intractably compounded when the dataset is massive. This is an issue of interest not just to statisticians but also to researchers from several fields such as the computer and information sciences or computational biology and bioinformatics. My multi-staged iterative approach (Maitra, 2001) was the first formal statistical approach to this difficult problem and also marked my first foray into the area. This diversified with the development, with a recent Ph.D. graduate (Maitra and Ramler, 2009), of efficient methods for clustering in the presence of scatter (observations that are unlike any other), with application to identifying industrial facilities with similar kinds of mercury releases for use in the potential development of more targeted and effective policies in combating toxic mercury effluence. We have also, in a paper under revision, extended these methods for sphered and directional data. Separately, we (Maitra and Ramler, 2010) have developed the k -mean-directions algorithm

for fast clustering of datasets that have unit norm, with application to clustering time-course gene expression sequences and managing schedules of conference presentations at professional meetings, such as for the Joint Statistical Meetings (JSM). A former graduate research assistant and I (Chen and Maitra, 2011) also developed a computationally efficient method for clustering autoregressive time series data and applied it to the context of clustering mutual funds with the aim of helping individual and institutional investors build a diverse portfolio.

A major challenge in finding homogeneous groups in data is that of quantifying the strength of support in the dataset for these groups. Another recent graduate student and I have developed methods for assessing significance of identified groupings in terms of an universally understood measure, namely, the p -value. The result is a *quantitation map* which provides the researcher with a quantitative summary indicating the strength of support for more complicated models relative to simpler ones. Our recent work (Maitra, Melnykov and Lahiri, 2012) developed bootstrap approaches under semi-parametric clustering methodology, within the ambit of which lies the k -means clustering algorithm. We have applied this development in the context of color quantization of a digital image on different computer displays, obtaining both the *minimal* and *optimal* colors that are needed to adequately display an image. Another manuscript (Maitra and Melnykov, 2012), under invited revision for the *Journal of the Royal Statistical Society Series B*, has developed methodology for the case when the groups have an underlying parametric statistical model, and applied it to the problem of identifying voting blocs among the 100 Senators in the 109th United States Congress. Recently, we have extended the above methodology to the context of semi-supervised clustering *i.e.*, when some of the observations have labeled information. Significance quantification in clustering also has uses with regard to regression-style variable selection, which I plan to develop. It is my view that forward variable selection can provide a promising approach to developing practical methodology for clustering datasets that have a large number of variables, or when the coordinate information is available only incrementally, such as in the case of temporal financial and profile information. One of my next objectives is also to develop methodology for performing diagnostics, in terms of identifying outliers and influential observations in the context of clustering datasets.

As mentioned earlier, clustering is a challenging task needed in many different applications. There is a plethora of available methods, but most of them are empirical with no proper understanding of the conditions under which they work well or poorly. In order to facilitate this understanding, we (Maitra and Melnykov, 2010) recently developed an overlap measure between different components in finite mixture models and clustering, along with efficient simulation methods for the same. Such simulation methods make it possible for researchers to properly evaluate any clustering algorithm on datasets with varying levels of difficulty. A publicly available open source R package MIXSIM, and a more comprehensive open source C package CARP (Melnykov and Maitra, 2011) has also been developed.

The preceding research connects very well with my interests in developing new approaches to simulation and stochastic computation. In this context, a collaborator and I (Ellis and Maitra, 2007) developed an efficient two-stage rejection scheme for obtaining realizations from extreme regions in multiple dimensions. I have also continued my interest in developing parallel simulation and estimation approaches to Markov Processes and Markov Random Fields (MRF's). Specifically, I have been investigating the development of multi-grid methods for simulating such fields, through the introduction of parallel MRF's at coarser scales — which, because of their fewer coordinates, have greater mobility around the statespace. A challenge here is the amalgamation of the different scales while keeping the marginal distributions at each scale straight-forward. My recent Masters' graduate advisee (Adam Pintar) and I have developed such methods for some non-Gaussian MRFs and applied it to the practical problem of determining the order in which Plato wrote his treatises.

I have also expanded on my earlier interests in the development of quantitative inference tools to guide both research and clinical diagnosis in medical imaging. I recently derived a practical generalized cross-validation (GCV) strategy for selecting the optimal bandwidth of two-dimensional Positron Emission Tomography (PET) image reconstruction, using a rewrite of the spectral decomposition of symmetric multi-dimensional circulant matrices in real form. This is shown to improve the quality of reconstructed clinical PET images. I have also collaborated in the development of methodology for more accurate noise estimation (Maitra and Faden, 2009) in Magnetic Resonance Images (MRI) and also (Maitra and Riddles, 2010) in their prediction (synthetic MRI) at unobserved settings from three acquired images. This last development means that images that can not be easily acquired because of technological or patient limitations can now be predicted quite accurately from images obtained at other easier-to-acquire settings. In both cases, variability assessment is also possible as extensions of the estimation and prediction methodology.

My research interests in imaging also include the development of functional Magnetic Resonance Imaging (fMRI) as a tool to understand cognition, *i.e.* to locate and understand areas of activation in the human brain in response to different tasks or stimuli, and to eventually use this information to detect anomalies in pathological cases. The statistical challenges posed here are both classical and modern: for instance, identification of voxels showing significant activation is the age-old multiple testing problem severely compounded by the fact that only at most 2-3% of the almost 100,000 voxels imaging the brain volume are expected to be activated. Further, the distribution of the noise at each voxel is not Gaussian as routinely assumed, but is Ricean. Thus we have, at each voxel, correlated time series that are marginally Rice distributed. I have, along with a dissertation advisee (Daniel Adrian) and another collaborator (Daniel Rowe), defined stationarity in Ricean time series and also developed methods to estimate the model parameters. The Rice distribution of these time series datasets, however, arises because the data used are really magnitudes of complex-valued Gaussian time series. The phase angle of the complex Gaussians are traditionally discarded in fMRI, which we have shown is not a good strategy. In doing so, we have also developed methodology for (complex-valued) bivariate Gaussian time series.

In a different context, but still within fMRI, my recent work (Bhattacharya and Maitra, 2011) provided a nonparametric approach to modeling nonstationarity and was used to understand the “attention control network” neuronal mechanism by means of which the brain sifts out task-relevant information from that unrelated to the task. This has provided a deeper understanding of the brain’s processing of the shape of an object, prompting a reviewer for the manuscript to note his thought that it “contained some important insights and developments.”

I also have a long-term goal of understanding the degenerative effects of neurological and other disorders with a view to facilitating improved patient care and therapy. Consequently, I have been involved in the development of methodology to assess the reliability of the observed activation. I have very recently (Maitra, 2009; Maitra, 2010) introduced methods to redefine and to generalize and thus obtain more consistent summary measures of identified activation across different subjects and/or studies. This has provided the ability to identify anomalous fMRI studies with the goal of improving the quality of data (and thus of inference). In multi-subject studies, they can also be used to identify pathological cases and to understand differences in responses of different subjects. Of course, in addition to accurate assessment of reliability in activation detection, it is important to improve the accuracy to detect such activation itself in the first place. My recent funding proposal to the National Institutes of Health proposes to provide consistent and reliable activation detection by explicitly specifying *a priori* expectations of extent and location of such activation while also maintaining spatial context.

Still within the context of imaging, but beyond medical applications, I have very lately become inter-

ested in the development of statistical methodology for non-destructive evaluation (NDE) of materials using vibrothermography which is a technique used for detecting cracks in industrial, dental, and aerospace applications. Here, a sonic or ultrasonic energy pulse is applied to cause the unit under test to vibrate. If a crack exists in the unit, it is expected that the faces of the crack will rub against each other, resulting in a temperature increase near the crack. An infrared camera is used to measure the temperature change in a sequence of frames over time. The technology is in its infancy, with modest development in the analysis of images: most of the available methods ignore the time series of images, focusing only on the sole image deemed to be the one with the highest signal. I am interested in developing methodology for comprehensively assessing the entire sequence of frames. In the long run, the goal is to be able to identify cracks bigger than a certain size automatically, while also to evaluate the growth trajectory of smaller cracks from one inspection to the other.

Many of the research issues presented here arise because of the rapid advances in technology. My current and planned solutions are built on exploiting computational resources such as modern workstations, parallel virtual machines and supercomputers, distributed and cloud computing, each of which need to be harnessed both smartly and efficiently. Most importantly, I believe that the research challenges discussed here will only become more acute with technological advances and with the development of more efficient and automated data collection methods. Thus, the significance of my research objectives extend well beyond immediate needs and will potentially be relevant for a long time.

Teaching Objectives

The primary role of an university is to disseminate knowledge and to prepare well-equipped citizens for society. It does this by fostering analytical thinking and by preparing students for the many facets of an increasingly technologically-driven lifestyle. Statistical education is no different, whether in the undergraduate classroom, the graduate level or in a mentoring context. This principle has always been the cornerstone of my teaching philosophies at both the UMBC and at Iowa State University, even though the emphasis has been modulated for each student as per his/her exact needs, educational background and goals. I consider teaching to be a very fulfilling and learning experience for me and believe that I have myself grown as an educator over the years.

I believe very strongly that undergraduate and service education is the foundation for a student's future in any discipline. At the same time, graduate classes develop in our students the expertise to practice statistics and to extend the frontiers of research. During my appointments, I have taught introductory undergraduate classes in statistics for students majoring in the social, engineering, pure and life sciences. While doing so, I have tried to relate my classes to my students' major disciplines with examples from day-to-day life and relevant applications. I have also taught upper-level undergraduate classes in applied statistics and in statistical software in addition to designing an undergraduate special topics class in data mining and in statistical computing. I have also developed undergraduate research experience classes for Honors freshmen interested in statistical research at Iowa State in Spring 2006. At the graduate level, I have taught core classes in applied statistics and designed and taught more advanced courses in multivariate statistics, nonlinear models, statistical computing and spatial statistics and image analysis. In many of my undergraduate and graduate classes, I include real-world data analysis project assignments, with professional-grade reporting requirements that emphasize clear communication, consistent with our society's need for promoting writing across curricula. In my advanced-level graduate classes, the assigned projects are intended to provide students with a full research experience: indeed, some of them have eventually been developed further and submitted to scientific journals or presented at professional meetings. Students have been very appreciative, with

complimentary comments often many years after the class is over! In addition, the Department of Statistics at Iowa State has a mechanism whereby a senior professor randomly visits another faculty member's classroom for two class hours every two years to provide detailed feedback to him/her and the chair. In this context, Professors Max Morris, Bill Meeker and Fred Lorenz have visited my classes in multivariate statistics and statistical computing in Spring 2004, 2006 and 2010 and Fall 2011 and independently provided very positive and helpful peer evaluations.

My teaching responsibilities have provided me with the experience to contribute effectively to both the development of curricula in different areas of statistics (*viz.* statistical computing, data mining, spatial statistics and image analysis). I was very closely involved with the successful effort to develop Maryland's first undergraduate statistics major and minor programs at UMBC. More recently, I was charged with leading the effort to recast the graduate curriculum in statistical computing at Iowa State. The new sequence has now completed several cycles and its content has been viewed very enthusiastically, both by faculty and students. Indeed, my biennial course on Advanced Statistical Computing (Stat 690E/680) has had to be run annually for several years because of student and faculty interest. I have also been very active in the mentoring of graduate students, postdoctoral researchers and junior faculty. I was the dissertation adviser for five students and am currently supervising one student. In addition, I have supervised the research assistantship of eight graduate students, in addition to directing the "creative component" research of seven M. S. students in the Iowa State statistics program. Two of my mentored students won American Statistical Association (ASA) Student Paper awards in Statistical Learning and Statistical Computing, respectively. I also volunteered (Spring 2005, 2009 and 2010) to mentor and teach additional classes (HONS 290) for incoming Honors freshmen and to provide them with hands-on research experience in applied problem areas in statistics. I have mentored undergraduate and graduate students as part of the National Science Foundation's (NSF) VIGRE-funded project at Iowa State – in this context, I developed and continue to lead the statistical computing working group component which grew out of the VIGRE program. I was a senior personnel in the department's successful Research and Training Graduates (RTG) funding request to the NSF for supporting graduate students. These proved to be very exciting learning experiences for me, as I was able to combine teaching and mentoring with the thrill of extending research frontiers while also obtaining experience in the writing of department-wide training grant proposals and in the supervision of funded research. I believe that these experiences will be very useful launching pads for the challenges ahead. Additionally, my research objectives gel well with the requirements of new graduates in an increasingly technological world. For instance, a substantial part of modern research in statistics involves scientific computation. Therefore, we need to continuously fine-tune and update both undergraduate and graduate instruction to keep abreast of such trends. This can be augmented through the development of suitably-tailored classes as well as through hands-on research experience projects. I consider myself well-equipped and experienced to further develop and re-design such classes which use both traditional and modern methods of instruction to address changing future needs of the department and the university.

Service and Professional Practice

An integral part of my professional responsibilities has been service to the discipline, the department, the university and the community. I have provided consulting services on demand from members of the basic sciences, computer science, engineering, biomedical and public health communities. Some of it has involved the design and analysis of clinical trials experiments, while others have dealt with more substantive research questions on clustering, imaging and stochastic computation. I have reviewed manuscripts submitted by my peers in several journals as well as proposals submitted for funding

to both the National Science Foundation (NSF) and the National Science and Engineering Research Council (NSERC) of Canada. I am currently the American Statistical Association (ASA) Executive Editor for *Statistics Surveys* – a position that I was appointed to, after being ASA Associate Editor for the same journal for three years. The ASA Executive Editor is appointed by the ASA with primary responsibility to appoint the Associate Editors (AE), to assign submitted manuscripts to the appropriate AEs to begin the review process. I am also responsible for arriving at a final decision on the review of a manuscript and to address author grievances. I am also an Associate Editor for *Sankhyā* Series A. Additionally, I am a Founder-member and Publications Officer of the new ASA Section on Statistical Imaging. I previously served on the Executive Committee of the American Statistical Association's (ASA) Section on Statistical Computing (1997–2005) and was its Continuing Education Liaison (1997-2001) as well as the Editor of the Statistical Computing and Graphics newsletter (2002-05). Additionally, I was elected President (2000-01, 2001-02) and Vice-President (1998-99, 1999-2000) of the Maryland Chapter of the ASA. I am also a representative of the International Indian Statistical Association (IISA) in the Program Committee of the 2012 Joint Statistical Meetings.

At the departmental level, I was a member of the Department Computer Advisory Committee (2003–11), and chaired it for two terms. This committee advises the department chair in supervising the systems administration staff and graduate students. Computing being the life-blood of our profession, it also develops a vision for the computing needs and directions of the department. I chaired the department's Journal Ratings Committee in 2006. This one-time committee was tasked with deriving an objective ranking of journals published in the discipline, with the goal of mentoring graduate students and junior faculty members by providing them with a list of premier and outstanding statistics journals where they could target their research and publications record in order to have maximum impact. In addition, I have also been a member of the Department's Seminars, Search, Strategic Planning, Journal Ratings, Graduate, Undergraduate and the Awards Committees. I have also played an active role in the recruitment of undergraduate and graduate students. As indicated earlier, I have been involved in efforts to develop programs and curricula. I believe that my experiences here will come in useful as new programs and courses are needed to be developed and existing ones modified. I coordinated the department's successful effort to get NSF funding under the SCREMS program. Additionally, at the university level, I am currently the department's elected representative in the University's Faculty Senate (2009-12) and also served on the Senate's Appeals Committee (2010–2011). At the community level, I have participated in the ASA's Adopt-a-School program which enables professional statisticians to help schools in their efforts to bring quantitative literacy into their classrooms. Thus I believe that my active service to the department, university, community and profession has been an enriching experience and has prepared me for further challenges ahead.

In conclusion, I believe that my research, teaching and service goals are very important and relevant to the needs of modern society. Indeed, they sit well with the cross-disciplinary flavor of modern statistical research and are in tune with the professional interests of researchers across many disciplines. These are very exciting times for statisticians and my past professional experiences and achievements provide me with the confidence that I can contribute effectively in the onward evolution of society, while also developing further as a valued scholar, teacher and member of the community.