Research Statement

Mark H. Hansen

Today, almost every aspect of our lives can be "rendered" digitally. Advances in data collection technologies have made commonplace continuous, high-resolution measurements of our physical environment (weather patterns, seismic events). Equally open to observation are our routine movements through and interactions with our physical surroundings (automobile and air traffic, large-scale land use). In computer-mediated settings, our activities either depend crucially on or consist entirely of complex digital data (financial transactions, Web site and Internet usage). As a reflection of the diversity and variety of the systems under study, these data-based descriptions of our daily lives tend to be massive in size, dynamic in character, and replete with rich structures. The advent of these enormous repositories of digital information presents us with an interesting challenge: how can we represent and interpret such complex, abstract and socially important data?

Since joining Bell Laboratories, I have had the opportunity to explore these issues through several significant collaborations. There are very few disciplinary barriers within the Labs and I have had the privilege to work with researchers from a variety of different fields, including physics, numerical analysis, information theory, signal processing, computer science, and most recently media arts. Below I have selected several research areas in which I have developed techniques for analyzing and modeling large, complex data sets.

Minimum Description Length

By viewing statistical modeling as a means of generating *descriptions* of observed data, the Principle of Minimum Description Length (MDL) suggests discriminating between competing models based on the *complexity* of each description. This approach began with Kolmogorov's theory of algorithmic complexity, matured in the literature on information theory, and has recently received renewed interest within the statistics community. In the substantial review article Hansen and Yu (2001), we explore both the practical as well as the theoretical aspects of MDL as a tool for model selection, emphasizing the rich connections between information theory and statistics. At the boundary between these two disciplines, we find many interesting interpretations of popular frequentist and Bayesian procedures. We illustrate that MDL provides an objective umbrella under which rather disparate approaches to statistical modeling can co-exist and be compared. In Hansen and Yu (1999a) and Hansen and Yu (2001), for example, we derive MDL criteria that adapt to the problem under study and behave like either AIC or BIC depending on which is most appropriate.

In terms of new methodological work, we have applied the MDL framework to the problem of wavelet denoising and compression for natural images. After making a connection between thresholding and model selection, we derive an MDL criterion based on a Laplacian model for noiseless wavelet coefficients. In Hansen and Yu (1999b) and Hansen and Yu (2000) we show that this approach leads to an adaptive thresholding rule and demonstrate that while achieving mean squared error performance comparable with other

popular thresholding schemes, the MDL procedure tends to keep far fewer coefficients. From this property, we demonstrate that our method is an excellent tool for simultaneous denoising and compression. We make this claim precise by analyzing MDL thresholding in two optimality frameworks; one in which we measure rate and distortion based on quantized coefficients and one in which we do not quantize, but instead record the rate simply as the number of non-zero coefficients. This approach is similar in spirit to the current work of Johnstone and Silverman on wavelet denoising.

Yu and I are in the beginning stages of a monograph on information theory and statistics which will build on our successful review paper. In addition to the basic results from coding theory (Shannon's Source Coding Theorem, Kraft's Inequality, universal source coding) and algorithmic complexity (Turing machines) that motivate MDL, we also plan to cover more exotic topics like context models and quantum information theory.

Multivariate Function Estimation

Polynomial splines and their tensor products are natural building blocks for constructing finite dimensional estimates of infinite dimensional main effects and low-order interactions, and the resulting ANOVA decompositions provide an insightful tool for data analysis. In order to synthesize the theoretical properties of these estimators, I developed the notion of an extended linear model (ELM). Many statistical problems of theoretical and practical importance can be effectively treated within this framework: Regression, generalized regression, polychotomous regression, hazard regression, censored regression, density estimation and conditional density estimation are all examples of ELMs. In my dissertation, I derived general results about the asymptotic properties of polynomial splines when applied in the context of an ELM; see Hansen (1994) and Stone, Hansen, Kooperberg and Truong (1997).

In addition, I established rates of converges for estimators that model interaction effects with non-tensor product, multivariate splines. While tensor-product splines take (hyper) rectangles as their underlying grid or mesh, many other spaces of smooth, piecewise polynomials are defined relative to triangles or simplices. These spaces have been very popular in the literature on numerical analysis and computer-aided geometric design, but have received little attention from statisticians. In Hansen, Kooperberg and Sardy (1998), we explore the practical aspects of modeling with triangulation-based splines and present a spatially adaptive, bivariate technique for regression and density estimation known as *Triograms*.

Many spline-based procedures depend on some kind of stepwise modeling to adapt the underlying mesh to features evident in the data. Increasing or reducing flexibility across regions of the input space usually translates into the simple addition or deletion of a single predictor (spline basis element) from an ELM. The fitting procedures tend to be *greedy* in the sense that basis elements are added sequentially, at each step choosing the one that creates the greatest boost in the likelihood. This is followed by a pruning phase where unnecessary terms are removed, at each stage eliminating the basis element that results in the smallest change in likelihood (see Stone et al., 1997). In Hansen and Kooperberg (2001), we evaluate various non-greedy approaches to this problem. Our focus was primarily on Bayesian schemes originally adapted for

generic model selection problems. The resulting stochastic methods depend on Markov chain Monte Carlo to explore different mesh configurations. In Hansen and Kooperberg (2001) we quantify the price paid for a simple, greedy search schemes, demonstrating that the loss is not too great for most practical applications.

The basics of spline-based modeling for a variety of ELMs, non-tensor product extensions, and Bayesian computational techniques are all covered in an upcoming monograph, *Statistical Modeling with Spline Functions: Methods and Theory* by Hansen, Huang, Kooperberg, Stone and Truong (2002).

Modeling Web Usage

Current software offerings aimed at analyzing Web usage statistics are typically little more than accounting packages that track the number of times users request each page on a site and present plots of access counts against time or perhaps the geographic location of the users. In two patents Hansen and Sweldens (2001a, granted; 2001b, pending), we designed new analysis techniques and supporting visualization tools to help Webmasters and content providers better understand how visitors are using their site. In designing our displays, we formed a very tight coupling between the content on the site (the Web pages and other documents) and the statistical analysis of usage patterns. Our *data browser* is an effective platform for data analysis and usability assessment. Based on exploratory analyses with these tools, we have gone on to formulate models that describe how people browse the Web. In Sen and Hansen (2001) we discuss several such models. In this paper, we restricted our attention to user navigation patterns within a single site, and explored how well we can predict the pages a user will request given their previous activities on the site. If the prediction is reasonably accurate, we might consider "prefetching" pages before the visitor requests them, ensuring that the content is immediately available. A more conservative use for such predictions would be to simply update the freshness records in a proxy or network cache, eliminating unnecessary `If-Modified-Since` requests. Using data from the Web site for the Computing and Mathematical Sciences Research Division of Lucent Technologies (`cm.bell-labs.com`) we first evaluated the predictive performance of low-order Markov models for user requests. We next considered mixtures of first-order Markov models, achieving a kind of clustering of Web pages on the site. This approach performed well, while significantly reducing the space required to store the model. Next, we explored a Bayesian approach using a Dirichlet prior on the collection of links available to a user at each stage in their travels through the site. Finally, we extended this model to incorporate user-level profiling based on navigation patterns. Early experiments with variable length Markov chains (known as context models in information theory) suggest that these models will provide an even better fit to the observed navigation data.

Moving beyond predictive models, in Hansen and Shriver (2001) and patents August, Hansen, Lee, and Shriver (filed 2000) and Hansen and Shriver (filed 2001), we created tools to skim large quantities of user navigation data to improve Web services. We took as our example Web searching. After an analysis of the searching task, we developed a technique for extracting "search sessions" from proxy log data. In this case, rather than observe the activities of all visitors coming to a fixed Web site, we see all the sites accessed by each user who is connected to the Web through a given proxy. Formally, a search session is

a collection of query terms together with the Web pages a user visited while searching. With these search sessions, we develop navigation-based models associating queries with Web content, deriving a kind of automatic categorization that is derived from subscriber activities. We also propose variants of the classical EM algorithm for training these models. In the end, we developed a system that dramatically reduces the number of items a user has to examine when searching.

In ongoing projects, we consider the concepts mentioned above (usability assessment, profiling, prediction and services enhancement) in a wireless setting. It is widely acknowledged that many wireless services are so poorly designed that they are essentially unusable. WAP (wireless application protocol) services are notoriously bad in this regard, often forcing wireless subscribers to contend with interfaces that abuse even existing non-mobile design principles. To be fair, application designers have little actual subscriber data for usability assessment because many of the platforms deployed by WSPs (wireless service providers) do not adequately track subscriber activities on their networks. Improved event logging is a critical first step to understanding subscriber behavior and evaluating usability. Despite these current deficiencies, the industry is expecting tremendous gains in wireless services. It is believed that subscribers will come to rely more heavily on smart, mobile devices, and that personalization is a critical component in creating new wireless services. To make this vision a reality, application designers need a more complete view of their subscribers. As wireless networks expand to carry new location-based applications and financial transactions, the WSPs will be collecting unprecedented amounts of subscriber data. How can application providers make use of these data? What tools should the WSP offer to help enable personalization and create new services?

Together with Shriver, I have been researching the data analysis infrastructure required for a WSP to offer usability assessment for existing services and to create new personalized services based on models of subscriber behavior. This infrastructure will consist of software and algorithms for both event logging as well as for building, managing and applying subscriber-based profiles. To date, we have begun working with the Mobile Consumer Lab at Boston University (BU), and are negotiating data from a wireless search engine, PinPoint.

Quality Improvement in Semiconductor Manufacturing

Semiconductor manufacturing is a complex process involving hundreds of separate steps and lasting up to two months. Several hundred circuits are fabricated simultaneously on a wafer, and the wafers are themselves processed in groups called lots. A large manufacturing plant typically processes hundreds of lots a week. Unfortunately, this volume precludes engineers from exploiting information from all the available data source for process improvement. Shortly after joining Bell Laboratories, I helped foster a successful collaboration between members of research and product and process engineers at Lucent's (then AT&T's) fabrication sites. Initially, my work focused on modeling the significant spatial variation intrinsic to many semiconductor fabrication processes and evident in almost all of the measurements and tests performed on the wafers as they passed through the production line. While these spatial effects were widely acknowledged and in fact motivated a variety of time consuming, ad hoc visual techniques for quality improvement, very

little formal work had been done to systematically mine these data for clues about failing processes.

In Hansen and Nair (1996), Friedman, Hansen, Nair and James (1997), and Hansen, Nair and Friedman (1997, 1999), we map out a statistical strategy for routinely identifying significant patterns or clusters in post-production electrical tests and relating them to critical process information. The formulation, evaluation and ultimate implementation of these ideas would not have been possible without the extensible computing environment we developed for visualizing, manipulating and modeling mapped wafer data. This software platform is described in Hansen and James (1995), Hansen, James and Pregibon (1995), and Hansen and James (1997). The quality of our design is apparent in the ease with which we could incorporate new data sources and new types of measurements. The latest such addition was a set of tools for modeling optical inspection data (scans of a wafer's surface that find particles) and identifying areas of the process responsible for excess contamination (see Patterson, Hansen and Henry, 1999; Patterson and Hansen, 2000). Through a series of workshops held in the factories, we popularized our methods and uncovered new, interesting problems. Prior to the spinoff of Lucent's microelectronics division, our techniques had become an integral part of the routine data analysis carried out by engineers in the manufacturing facilities, and were estimated to save Lucent $15M per year.

In Nair, Hansen and Shi (2000) we discuss some of the challenges in developing statistical methods for quality, process and productivity improvement in advanced manufacturing settings. Much of the inspiration for this paper came from our experience with semiconductor fabrication. Interestingly, our work in microelectronics also provided motivation for a significant effort in designing generic computing tools for technology transfer. Chambers, Hansen, James and Temple Lang (1998) describes how this work evolved.

Sound as a Tool for Data Analysis

As part of its long-standing interest in breaking down the boundaries separating disciplinary genres, the Brooklyn Academy of Music developed a year-long artist-in-residency program in conjunction with Lucent Technologies. The Arts in Multimedia (AIM) project paired artists working in new media and communications technology with scientists at Bell Laboratories. In 2000, sound artist Ben Rubin (EAR Studio, New York City) and I received an AIM award for our proposal entitled *Ear to the Ground* (ETG). As a collaboration between a statistician and an artist, we are looking to create sound representations that further the arts of data analysis, discovery, and expression. We plan to establish a series of listening posts, points in the physical world or on the Web, where people can engage large, often abstract data streams, bringing an awareness and understanding of these data to the general public. By creating these installations, we also hope to establish general principles for the public display of complex data.

ETG's first listening post provided a sonification[1] of the traffic on a Web site. The richness of this representation is controlled by the number of visitors accessing *detailed* content (Web pages or other documents). The presence of each active visitor contributes to the loudness and tonal balance of a low-register

[1]The term "sonify" refers to the direct mechanism by which data re rendered audible. A "sonification" is the output of such an exercise.

drone. Each major section of the Web site is assigned a different pitch, the relative volume of which is increased as more visitors browse content from that area. Requests for more informative content deep in the site are represented as higher-pitched pulsing tones: the faster the pulses, the more people are accessing that area, and the higher the pitch, the more detailed the content. This initial work was featured in the online editions of the New York Times and Science News and is described in Hansen and Rubin (2000) and Hansen and Rubin (2001). A sample (Rubin and Hansen, 2001) also appears on a CD entitled "State of the Union," produced by Elliot Sharpe.

ETG's second project titled Listening Post was an attempt to characterize the "collective voice of the Internet." Over the last few years, the Web has become a massive communications channel: Conservatively, 35,000 sites advertise some form of chat or online forum, and recent IRC (Internet Relay Chat) statistics suggest a user pool of over half a million people. Through Listening Post, we offered the public a view of this vast world of online activity, distilling content from tens of thousands of online exchanges in real time and revealing the patterns and rhythms of people communicating with each other. At the center of this uniquely designed space was a large array (7' tall by 10' wide) of 110 vacuum fluorescent displays (VFDs), each screen holding up to four lines of 20 characters. This grid displayed portions of online conversations, and cycled through several modalities in an attempt to convey scale (the number and diversity of Web discussions) and immediacy (that the displayed pieces of conversations are generated in realtime), while preserving the character (the individual exchanges mediated through various chat programs) and content (common topics of discussion and their frequency) of the underlying sources. In addition to the grid of VFDs, Listening Post also made use of a multi-layered soundscape projected through a series of speakers spread throughout the installation space. Lucent's text-to-speech engine, the Articulator, was a large component of the audio, giving voice to many of the chat fragments displayed on the VFDs.

The technical challenges implied here are considerable; from scalable monitoring agents that continually recognize and cull new content on the Web, to statistical natural language processing and dynamic clustering schemes that allowed us to track topics and extract representative phrases. Toward this end, we constructed a network of computers that continually monitor 10,000 chat rooms and bulletin boards. To be efficient, the monitoring agents adapted their polling rates to the activity levels of the sources. As discussions took place, the content was categorized into topics based on (learned) key words, the participant's previous posts, and other recent or concurrent discussions. For future installations, we are also considering covariates drawn from the Web site sponsoring the forum, as well as live news feeds. This categorization is by necessity dynamic, changing over the course of a day, and is highly influenced by current events. The design of the Listening Post is described more fully in Hansen and Rubin (2001,2002).

Listening Post was installed as part of the Brooklyn Academy of Music's 2001 Next Wave Festival and was open to the public from December 6 – 20, 2001 (see `www.bam.org`). Additional funding for this piece has been supplied by a grant from the Rockefeller Foundation. Other Listening Post milestones include an invited presentation at The Kitchen, a performance space in NYC that promotes interdisciplinary work (`www.thekitchen.org`); an invited talk for the National Academy of Science's

Committee on Information Technology and Creativity; participation in an invited roundtable organized by Creative Time & Thundergulch (two well-known arts organizations in NYC – `www.creativetime.org` and `www.thundergulch.org`) called "Blur: New Creative Practices Within Developing Technologies" that dealt with how artists and technologists are approaching living social systems; and an invited presentation at ArtSci2001, the third international symposium organized by Art and Science Collaboration, Inc. (`www.asci.org`). Finally, this work was featured in the New York Times, NPR's Studio 360, and the January 2002 issue of Art Forum magazine.

Last year, I also started to explore the use of sound as a tool in support of more traditional data analysis activities. In Hermann, Hansen and Ritter (2001), we develop an approach for creating real-time sonifications of Markov chain Monte Carlo (McMC) simulations. Our method consists of several auditory streams which provide information about the behavior of the Markov chain and the target density p. Using a technique known as model-based sonification, we let p represent a gravitational field and consider the trajectories of particles thrown into the domain of p at a given position and with some kinetic energy. At each step in the McMC simulation, we emit a particle from the chain's current position and track its motion, listening to a smoothed version of its kinetic energy. If the chain is close to a mode, the particle will undergo damped harmonic oscillations as it moves through an approximately parabolic potential. This gives rise to a pitched audio signal, the spectrum of which corresponds to the eigenvalues of the Hessian matrix at the mode. From this, we get information about the mode shape. As the chain moves from one state to the next, we hear a sequence of pitches that offers insight into the mode structure of p, as well as the mixing time of the chain. We have also found our sonification to be an effective means for understanding the structure of high-dimensional data sets. In this case, we take p to be a nonparametric density estimate obtained from the data. Recently, we have been exploring the connections between our approach and the so-called Hybrid Monte Carlo methods from statistical mechanics.