Statement of Research Interests and Plans

Jaimyoung Kwon
UC Berkeley

My general research interests are twofold, one application and another theoretical, which is clear from the title of my thesis (Kwon, 2000). More specifically, they are application of statistics to transportation science and bioinformatics, and extension of calculus of efficiency to general non-IID cases.

Application of statistics to transportation science

On the application side, my main focus has been applying statistics to transportation science. As a graduate student researcher and then as a postdoctoral researcher, I have been involved in a research subgroup of Partners for Advanced Transit and Highways (PATH; see http://www.path.berkeley.edu), an interdisciplinary group of statisticians, electrical and civil engineers and computer scientists working in the field. Datasets considered in the field are usually extremely large and complex, posing many computational and conceptual challenges for statisticians.

Loop data

The most common and basic data in transportation science is loop detector measurements of traffic flow. According to PeMS website (see below), in LA County alone, about 1770 loops are installed on highways, producing flow (vehicle counts) and occupancy (a proxy for the degree of congestion) every 30 seconds. The dimensionality of the data is astronomical and the traffic flow phenomenon is extremely complex. It is fairly recent that such data has become publicly available and we possess enough computing power to analyze the data. Our group is working on a Freeway Performance Measurement Project (PeMS; see http://transacct.eecs.berkeley.edu), a system that collects historical and real-time freeway data from freeways in the California in order to compute freeway performance measures. It consists of a web inventory and automated analysis/reporting tools to streamline collection, processing and analysis of the loop detector data.

I have been working on applying statistics to analyze the data and improve background algorithms of PeMS website. The specific research programs include:

- Sensor malfunction detection based on censor-to-sensor correlation (Kwon, Bickel and Rice, 2002),
- Imputation of bad loops using graphical models (work in progress),
- Summary and visualization of the high dimensional spatio-temporal data (Kwon, Coifman and Bickel, 2000),
- Travel time prediction (Kwon, Coifman and Bickel, 2000 and Bickel et al. 2001),
- Phenomenological modeling of congestion dynamics using coupled hidden Markov models (Kwon and Murphy, 2000)
- Better estimation of vehicle speeds from ‘single-loop detectors’ data, improving works of Dailey (1999) and Jia et al. (2001a) (Work in progress)

Since loop detectors are already installed in most urban freeways in the US, the data inventory is ever-growing and more effort will be made to use the data for Advanced Traveler Information Systems (ATIS) and Intelligent Transportation Systems (ITS). Already, some efforts have been made to use the data for metering (Chen and Varaiya, 2000), capacity analysis (Jia et al., 2001b) or for data mining (Chen et al., 2001) as well as the ATIS component (travel time prediction) of PeMS website. Innovative statistical methodology is crucial in developing state of the art algorithms for these problems.

Video data; computer model

Another interesting source for studying traffic flow is video data. We are acquiring video data on a 2.5-mile stretch of I-880 freeway (Coifman et al., 1999), which was processed by a computer vision algorithm (Coifman et al., 1998) that tracks vehicles to produce vehicle trajectories at 0.1-second time

resolution. The data is growing as more video data is collected and processed, becoming another large dataset to analyze. Currently, about 1,500 hours of video data have been collected.

It is an invaluable data source for examining microscopic and medium-scale driving behavior of drivers, which has rarely been studied rigorously. Again, the size of the data and the complexity of the phenomenon are quite challenging and statistics can contribute a lot to process, visualize and analyze the data. A few research areas I plan to consider include:

- Individual driver's behavior at micro-scale. It is essentially a huge longitudinal data with many subjects (hundreds of drivers) with variable observation time (during the vehicle is within camera's field of view, ranging from 5 seconds to 30 seconds). Longitudinal data analysis and/or mixture model can be applied. But, the size of the data can lead to a computational problem.
- Group behavior at medium scale: it is known that vehicles form 'platoons' while they travel on freeway. This medium-scale behavior is an interesting dynamic point process, which may require development of new stochastic process to model.
- Visualization of the video data. The processed data consists of trajectories, i.e. x and y coordinates over time, of many vehicles. Visualization of such high dimensional longitudinal data is extremely challenging and would require interactive graphics as well as animation. I expect the visualization techniques developed for this data are widely applicable for other high dimensional longitudinal data with many subjects.

Evaluation and calibration of complex computer models

CORSIM (CORridor SIMulation; FHWA, 1997) is a "computer model" that tries to mimic actual traffic flow by mimicking behaviors of individual vehicles in the virtual freeway environment. They are widely used for various practical traffic-engineering activities such as signal retiming, traffic impact studies, analysis of major traffic events, stadium operations, corridor traffic operations, and freeway incident impacts. They are also used to evaluate ITS technologies, such as real time traffic adaptive control, real time traveler information and route guidance, and network-wide dynamic traffic assignment.

The parameters of driver's microscopic behavior are key tuning parameters for microscopic simulation models like CORSIM. Improved estimation of those parameters from the video data will have important implications on tuning and calibration of the CORSIM and other microscopic traffic simulation models.

More generally, complex computer models are increasingly being developed to approximate real phenomena in various fields of science and policy. Many of those computer models are "cellular automata", which simulate complex phenomena by viewing whole populations as interacting "cells" or "automata" that follow a set of rules. Those rules can be derived from a theory, estimated from empirical data, or both. Evaluation (how good are the approximations?) and calibration (what can be done to improve them?) are crucial for sensible operation of such models. To articulate the roles statistical concepts and tools can play in these subjects, I would like to investigate the following:

- Developing better evaluation schemes for microscopic freeway traffic simulations, using the video data as the ground truth. Generalizing the evaluation schemes to a more general computer model context.
- The effect of more realistic tuning parameters on the performance of CORSIM. How much will the parameters of driving behavior validated by the video data improve the performance of CORSIM?

Statistical methodologies for complex dependencies in spatio-temporal data

Although the techniques developed for this traffic data are useful per se, some of them have potential applications to other areas. For example, the web of evidence model (Kwon, Bickel and Rice, 2002) which is a graphical model with a Bayesian flavor, developed for loop malfunction detection is an interesting model and may be useful for other areas where one has to detect malfunctions in an array of correlated sensors.

Very large datasets that have space and time dimension are now becoming more and more commonplace, arising from various fields including earth sciences like oceanography, climatology, and ecology. Chances are high that a set of frameworks and methodologies developed for the transportation

problem can be applied to problems in those fields. The coupled Hidden Markov Model (HMM) in Kwon and Murphy (2000) is an example of such efforts.

Relevant statistical methodologies for such applications of statistics to large spatio-temporal data with complex dependencies include time series, spatial statistics, HMM, graphical models and Bayesian analysis. I have a broad range of interest and am interested in such interdisciplinary communication and collaboration that will strengthen both areas.

Statistical genetics and bioinformatics

I have worked with biologists on statistical analyses of genomic and microarray experiment data. In particular, I worked on a new approach for a contemporary bioinformatics problem: finding yeast transcription factor binding sites from microarray data on the organism (Tavazoie et al., 1999). The idea was to use the soft membership probability instead of hard (zero-one) membership and incorporate the location specificity of known binding sites from database such as SCPD (Zhu and Zhang, 1999) to improve the pattern-scoring algorithm. Kielbasa et al. (2001) has recently utilized the latter idea independently.

Though it was hard to verify whether my proposal really improved the detection rate due to the lack of ground truth, I enjoyed learning genomics and seeing how statistics can solve interesting problem in genomics. Also, the size of the dataset (both yeast genome with about 6000 open reading frames (ORFs) and microarray dataset of those ORFs under hundreds of experimental conditions) was challenging as well.

There are various conceptual and computational challenges in the bioinformatics field that can be addressed well by rigorous application of statistics. I intend to contribute to the field by seeking out for collaboration with serious biologists.

Calculus of statistical efficiency in non-IID cases

I have been working on generalizations of calculus of statistical efficiency to non-IID situations (Bickel and Kwon, 2001). The theory has many potential applications as pointed out by the discussants of the paper. I have been applying the techniques to show efficiency of certain estimators under various settings, especially Markov chain time series (Kwon, 2000). Currently, I am considering a few nonparametric regression problems where the methodology can be applied to show efficiency of candidate estimators.

Summary

In addition to theoretical research programs on calculus of statistical efficiency in non-IID cases and nonparametric estimation in time series, I have a wide range of research programs for application of statistics to transportation science and genetics. Analysis of large and complex dataset has been my specialty and interest, whether it comes with time and space dimension (transportation) or without (genomics). I wish to investigate statistical analysis of such data further by employing methodologies from graphical models, statistics for time series, spatial statistics and Bayesian approaches. I also enjoyed working closely with scientists and engineers in various fields including transportation, biology, or geology (Kwon et al, 2001), applying various statistical techniques like bootstrap and modern regressions as well as the previous methods to answer questions arising from those fields. Such interdisciplinary collaboration has always been exciting for me and I want to pursue it further.

November 29, 2001.

Reference:

1. Bickel, P.J. and Kwon, J. (2001). “Inference for Semiparametric Models: Some Current Frontiers (with Discussions),” *Statistica Sinica* Vol. 11, No. 4, pp. 863-960.
2. Bickel, P.J., Chen, C., Kwon, J., Rice, J., Varaiya, P. and van Zwet, E. (2001). “Traffic Flow on a Freeway Network,” to appear in the *Proceeding of MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, California, March 19-29, 2001, Springer.
3. Chen, C., Petty, K., Skabardonis, A., Varaiya, P. and Jia, Z. (2001). “Freeway Performance Measurement System: Mining Loop Detector Data,” *80th Annual Meeting of the Transportation Research Board*, Washington, D.C., Jan 2001.
4. Chen, P. and Varaiya, P. (2000). “Optimal Ramp Metering Policy”. (Work in progress)

5. Coifman, B., Zhang, X., Lyddy, D., Skabardonis, A. (1999). “The Berkeley Highway Laboratory-Building on the I-880 Field Experiment.” (Submitted for publication).
6. Coifman, B., Beymer, D., McLauchlan, P., and Malik, J. (1998). “A Real-Time Computer Vision System for Vehicle Tracking and Traffic Surveillance”, *Transportation Research: Part C*, vol 6, no 4, 1998, pp 271-288.
7. Dailey, D.J. (1999). “A statistical algorithm for estimating speed from single loop volume and occupancy measurements,” *Transportation Research B*, 33B(5): 313-22, June 1999.
8. FHWA (1997). *CORSIM User's Manual*, U.S. Department of Transportation.
9. Jia, Z., Chen, C., Coifman, B. and Varaiya, P. (2001a). “The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors,” Submitted to IEEE 4th International ITS Conference.
10. Jia, Z., Varaiya, P., Chen, C., Petty, K. and Skabardonis, A. (2001b). “Maximum throughput in LA freeways occurs at 60 mph” (work in progress).
11. Kielbasa, M., Korbel, J.O., Beule, D., Schuchhardt, J. and Herzel, H. (2001). “Combining frequency and positional information to predict transcription factor binding sites,” *Bioinformatics*, 17, 1019-1026.
12. Kwon, J. “Calculus of Statistical Efficiency in a General Setting; Kernel Plug-in Estimation for Markov Chains; Hidden Markov Modeling of Freeway Traffic,” Ph.D. Dissertation, UC Berkeley.
13. Kwon, J., Bickel, P. and Rice, J. “The Web of Evidence: Detecting Malfunctions in an Array of Correlated Sensors.” In preparation, expected January 2002.
14. Kwon, J., Coifman, B., and Bickel, P., (2000). “Day-to-Day Travel Time Trends and Travel Time Prediction from Loop Detector Data,” *Transportation Research Record* no. 1717, Transportation Research Board, pp. 120-129.
15. Kwon, J., Min, K., Bickel, P. J. and Renne, P. R. (in press). “Statistical methods for jointly estimating decay constant of K40 and age of a dating standard,” *Mathematical Geology*.
16. Kwon, J. and Murphy, K. (2000) “Modeling Freeway Traffic Using Coupled Hidden Markov Models,” Technical Report available at http://www.stat.berkeley.edu/~kwon.
17. Petty, K., Bickel, P. Kwon, J. and Rice, J. (in press). “A New Methodology for Evaluating Incident Detection Algorithms,” *Transportation Research C: Methods*.
18. PeMS Group: Freeway Performance Measurement Project Website http://transacct.eecs.berkeley.edu
19. Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church (1999). “Systematic determination of genetic network architecture,” *Nature Genetics* 22, 281-285.
20. Zhu, J. and Zhang, M.Q. (1999). “SCPD: A Promoter Database of Yeast Saccharomyces cerevisiae,” *Bioinformatics*, 15:607-611.