Mathematical Conversations

David Siegmund: Change-point, a consequential analysis >>>

David Siegmund

Interview of David Siegmund by Y.K. Leong (matlyk@nus.edu.sg)

David O. Siegmund is widely acclaimed for his fundamental contributions to the theory of optimal stopping time in sequential analysis and for his recent work on the application of analysis to genomics. He is well-known for his philosophical delight and mathematical ability in commuting between the theoretical heights of probability theory and the murky depths of statistical applications.

He taught for about 8 years at Columbia University, where he obtained his doctorate under the supervision of Herbert Robbins. Since 1976, he has been at Stanford University, where he was Chair of the Statistics Department twice, served as Associate Dean of the School of Humanities and Sciences and is now the John T. and Sigrid Banks Professor. He has been a visitor to The Hebrew University, University of Heidelberg, University of Cambridge and Oxford University. He was at NUS in 2005 as the first Saw Swee Hock Professor of Statistics.

He has been invited to give lectures at major scientific meetings; in particular, the Wald Lectures, Hotelling Lectures at the University of North Carolina, Taiwan National Science Council Lecture, and Bahadur Lectures at the University of Chicago. Among the many awards he received are the Guggenheim Fellowship, Humboldt Prize, Wilks Medal and membership of the American Academy of Arts and Sciences and of the National Academy of Sciences of USA. He has served extensively on professional committees in the US. He has also been on the editorial boards of leading journals, such as the Annals of Statistics and the Annals of Probability. He was president of the Bernoulli Society and of the Institute of Mathematical Statistics. His numerous papers deal with theoretical questions in probability theory as well as concrete applications concerning clinical trials and gene mapping. He wrote two books (the first jointly with Y.S. Chow and H. Robbins) which are now classics in sequential analysis.

David Siegmund's long association with NUS dates back to the 1980s (as external examiner for the University of Singapore) and continues as a founding member of the Scientific Advisory Board of the University's Institute for Mathematical Sciences since 2001. When he visited the Department of Statistics and Applied Probability (DSAP) from October to December 2005, Y.K. Leong interviewed him at DSAP on behalf of *Imprints*. The following is an edited and revised transcript of this interview in which he talks passionately about his early attraction to mathematics, his subsequent search for the relevance of the mathematical sciences and a calling which he finds fascinating and challenging in theory and application. Here he also shares with us his rich experience in research and administration.

Imprints: Were you already fascinated by statistical mathematics in your school days? Were your school teachers instrumental in attracting you to statistics?

David Siegmund: The answer to the first part is clearly "no". In my school days, I had one mathematics teacher whom I liked very much, but at that time I was more interested in the foundations of mathematics. I found a book describing Cantor's set theory, the cardinality of infinite sets, the non-denumerability of the real numbers, etc. I thought that was a beautiful subject. I did have a university teacher who was instrumental in my attraction to statistics. In some sense, I became interested in statistics because I became disenchanted with the way mathematics in the 20th century had divorced itself from science. I took up an interest in this science and that science, shopping around, and at one point tried the social sciences. After deciding that none of these was exactly right for me, but with an interest in the social sciences, I was drawn to statistics as an area of mathematics closely related to the social sciences. Ironically, I have never done anything specifically related to the social sciences since then, but it did play a role in helping me find the field of statistics.

I: Were you more interested in applications than theory?

S: I've always been interested in theory. At heart, I would love to be a pure mathematician. At the same time I always wanted problems that seem to be related to some kind of applications, but they certainly don't have to be applied

Continued on page 8

Continued from page 7

problems in the sense that working applied statisticians would recognize them as applied problems.

I: How did you get interested in sequential analysis?

S: During my last year in the university, I took a course that involved reading Volume 1 of Feller's book on probability theory (at that time there was only one volume, now there are two), and I thought that the chapter on gambler's ruin was both fascinating and mysterious. The problems were fascinating, and while setting up difference equations was very natural, pulling solutions out of the air, as it seemed to me at that time, was very mysterious. In my first year at graduate school, I took a course in sequential analysis from Herbert Robbins and found the same problems were considered there from a completely different point of view. The methods of solution seemed more satisfying, and the connections to statistical applications added to my interest in the classical problem of gambler's ruin. Since then I have been interested in sequential analysis.

I: Was your PhD thesis on a topic in sequential analysis?

S: It was – on optimal stopping theory. One of the first things I read on my own during the first summer I was a graduate student was the chapter in Doob's book, *Stochastic Processes*, on martingale theory. I thought that it was the most beautiful mathematics I had seen up to that time, and it was naturally related to optimal stopping theory. Conceivably, I had the motivation from sequential analysis at the time but I don't recall. I think I just wanted to learn stochastic processes and that was one chapter that particularly appealed to me. Since my PhD thesis advisor, Herbert Robbins, was interested in optimal stopping theory, and it was naturally related to martingale theory, it was the subject for me.

I: That was at Columbia?

S: Yes, that was at Columbia.

I: Is Columbia near your home town?

S: No, I grew up in St Louis which is right in the middle of the United States. I started to think about Columbia because my wife was interested in going to the Columbia School of Social Work, probably the best known school of social work in the United States. When I mentioned this to Paul Minton, who advised me as an undergraduate, he became excited and said, "Oh, Herbert Robbins, now at Columbia, would be a wonderful advisor. He is very creative. You would love to work with him." So my wife's interests and my interests seem to coincide, and we went off to New York.

I: Robbins was originally a topologist?

S: He wrote his PhD thesis at Harvard in topology, but then before he really developed as a topologist, he was led during the war to problems of operations analysis. After the war he was invited to become a professor of statistics even though he had never taken a statistics course in his life.

I: In your scientific career, you have moved between Columbia and Stanford. What made you decide on Stanford as your eventual choice?

S: From a professional point of view, I found different advantages at Columbia and at Stanford, but my wife was an unequivocal spokesperson, on behalf of our children too, in favor of Stanford. I think she was completely correct – it is a much nicer place to live in than New York City. The scientific advantages became clear to me later on, though early in my career I liked very much to be in Columbia. But Columbia is not as strong a scientific university as Stanford is, and the statistical applications one naturally comes across in New York outside the university have to do with the financial community, the legal community and so forth. Those were interesting but I did not naturally gravitate to them the way I gravitate to some of the scientific things at Stanford. And Stanford's statistics department was larger and certainly, on average, a better department. So that seems to have been a good choice in the long run.

I: I believe you were at Columbia for quite a while.

S: I went there for three years as a graduate student and beginning assistant professor, with a one year hiatus at Purdue University, where Y.S. Chow was on the regular faculty and Robbins and Aryeh Dvoretzky were visitors. After two years at Stanford as an assistant professor, I went back to Columbia for seven years. But since 1976, I have been at Stanford.

I: Do you consider your work from sequential analysis to change-point analysis a natural development of your scientific interests? Could you tell us something about the origin of change-point analysis?

S: It was certainly a natural step. I really didn't know much about change-point analysis; but Bruce Macdonald, who headed the statistics section of the Office of Naval Research asked me to give a seminar in Washington, because he thought some of my research might have applications to change-point analysis. I went there with a few of my own thoughts, but in ignorance of the existing literature. Some of the questions asked by the audience and some of the references they mentioned made me aware that there was this field of change-point analysis. I realized that it was

Continued on page 9

Continued from page 8

indeed closely related to what I had been doing in sequential analysis and that it was quite interesting. In a sense, change-point analysis began with quality control at Bell Telephone Laboratories in the 1920s and 30s, but the real breakthrough, which ushered in the modern period, involved a couple of papers by E.S. Page, a British statistician, in 1954 and 1955 when he introduced the CUSUM test as a means of quality control. He didn't understand the relationship of the CUSUM test with the likelihood ratio test of statistics. That understanding came later, but since that period in the 1950s the subject has grown quite a bit. Initially, it was the result of the seminar questions that I didn't know the answers to, but then later the very rich theory and applications, that have held my interest.

I: You mentioned change-point analysis as a quality control thing. Was it empirically motivated?

S: Yes. The conceptual scheme is that we have a process, some kind of industrial process, that produces items in a complicated way that amounts to a black box. We can't look inside the black box to see if it is operating correctly. What we can do is to make measurements on the products to infer indirectly if it is operating correctly. The change-point philosophy was that you are careful in the beginning when you set things up, and the black box will initially operate correctly. Then after a while, someone gets careless or machinery wears out, and there is a change in the product, and you have to spot that change and then make adjustments to the system so that it starts operating correctly again.

I: Is there a theoretical foundation for this?

S: There certainly is a mathematical foundation. From the point of view of applications, there is always a debate whether a particular model is the best model that you can use. There are models where changes occur instantaneously by a discrete amount and others where changes occur gradually. There is a debate on which kinds of models are better. In spite of a certain level of implausibility, by and large the model that posits abrupt changes is very successful.

I: Do I understand that there are many change-point models?

S: Yes. There is no canonical problem. A problem has a certain structure to it but there is not a single mathematical formulation. In fact, I am sometimes at a loss for terminology. The term "change-point" is embedded in people's minds, but there are many problems with the same essential mathematical structure that don't really fit the change-point idea. So I sometimes use the phrase "change-point-like problems" to convey the idea that we are doing something related to change-point problems but it's not what you would automatically expect.

I: How do you choose the model to use when you are doing change-point analysis?

S: I don't think the answer is any different from any other statistical analysis. One typically starts with the simplest possible model that seems to capture some of the conceptual features of the problem, and then starts adding complications, sometimes called "bells and whistles", to make the model more satisfactory in a quantitative sense, although there is always the desire to keep the things as simple as possible conceptually. There's a famous statement of Einstein to the effect that a theory should be as simple as possible but no simpler. It's the same thing in choosing a model.

I: Do you know whether change-point analysis has been applied to data in the social sciences or even in the historical studies of cultures or linguistics?

S: There is a simple answer to the question, which is "yes", but I can't very effectively describe these applications. There are some in economics and finance, which in fact was the origin of some of the early applications of change-point analysis. In finance, for example, my colleague at Stanford, T.L. Lai has developed quite sophisticated change-point models that can lead to different investment strategies from time to time. I also occasionally get sent a paper or am asked to comment on a paper in the social sciences that has a change-point aspect to it. I usually forget these pretty quickly, so I don't really feel comfortable trying to discuss them in detail. But, for example, I do recall some research concerned with learning theory that asked the question whether learning, say simple skills in elementary school, should be thought of as something that proceeds by occasional dramatic improvements, where testing would indicate that someone hasn't learned anything but then seems to learn overnight, or alternatively that tomorrow we will be a little better than we are today and the next day we will again be slightly better. The learning theorist was trying to build a theory suggesting that progress appears to be rather abrupt, which would be consistent with a change-point model.

I: In history, for example, there are events which are marked by changes which can be thought of as change-points.

S: Right, there's certainly some of that motivation for applications in economics. People ask whether certain policy issues actually lead to changes in behavior or changes in economic conditions or whether certain external shocks to the system lead to a dramatic change or lead effectively to no changes at all. Conceptually that kind of issue has been a part of some economic thinking.

Continued on page 10

Continued from page 9

I: Is it possible to use change-point analysis to make history more quantitative?

S: I don't know. Historians try to use surveys and quantitative methods more and more. It would be interesting to know what kind of change-point models there might be. One problem that is an interesting conceptual application of change-points (and has an historical aspect to it) is the set of data examined by many people, which involves fatal accidents in British coal mines. For about 150 years, the British Coal Mining Board recorded accidents, and kept very clear records. Every accident that involved the deaths of more than 10 miners was recorded. During the period around 1890 there were royal commissions that studied the problem and made recommendations for how mining practices should be changed to make them safer. People naturally wanted to know whether this had an impact. Indeed, the rate of accidents dropped quite precipitously, or the average time between accidents increased quite sharply around 1891 - 1892, during the time that these activities took place. One presumes that this was a response to changes recommended by the commissions, which involve things like, if I recall, using a different kind of explosives, one that is less flammable, using water to wash down the interior of a mine, in particular, trying to get coal dust out of the atmosphere.

I: Is change-point analysis extensively used nowadays?

S: It is certainly widely used in the sense that you can find versions of change-point analysis in many, many different scientific contexts. Within those contexts, it's fairly specialized. For example, in drilling to find oil one wants to know something about the density of rocks through which one is drilling and in particular changes in density reflecting changes in the mineral composition of strata encountered during the drilling process. Change-point analysis of magnetic resonance image data is an approach to this problem that has a somewhat different flavor from most other applications I'm familiar with. Change-point analysis of DNA sequence data has recently become popular in some problems of molecular biology.

I: What about to evolutionary biology?

S: I guess there should be, but I've never looked at the data, and I don't know whether anybody has actually tried to formalize such a model. Certainly there is this ongoing debate about the hypothesis of Stephen Jay Gould of a punctuated equilibrium, that evolution doesn't proceed by small incremental changes as people more or less inferred from Darwin, but exists in a steady state without much in the way of changes and followed by a large number of changes occurring rather rapidly. I think this is a rather natural reaction to reflecting about the role of the environment in evolution, because we know that there are things like ice ages, meteors hitting the earth and volcanoes that have drastic impact on the environment leading to dramatic changes in, say, the average temperature of the earth and the seas. So it's natural to think that those changes, if they occur quickly, must lead to rapid changes in flora and fauna as well. But I don't know if anybody has actually tried to build a model and address the issues quantitatively. It would certainly be interesting, but it is also likely that the data are not sufficient, since this involves going a long way back in the history of the earth to find appropriate data. More modest questions of an evolutionary nature involve change-point analysis of DNA sequence data to identify, for example, places where mutations occur more frequently than the overall background rate.

I: Am I right to understand that the identification of a gene is a change-point problem in DNA analysis?

S: It certainly can be viewed that way. I would say it is helpful to view it that way, although most people involved in gene mapping, which is the area I'm primarily interested in now, do not share my view. I think they are missing something. With the advances made in technology that allow one to genotype markers closer and closer together, the change-point aspect of the problem will become more apparent. Historically, there were very few markers distributed across the genome. For the last ten years, in human genetics it has been customary to use on the order of 300 to 500 markers. Even at that level of resolution, the change-point aspect of the problem is not quite so apparent; but if the resolution should ever become what would be implied by having thousands of markers, which one can easily imagine, then the change-point viewpoint will increase in importance.

I: Is the problem of gene determination in the human genome completely solved?

S: No, it's one of those problems where progress seems very rapid, but then one realizes that there are still many more problems. With each step that we can take, we become more ambitious. Not so long ago one didn't try to map genes except for very simple diseases where there was one gene involved and the gene literally over-ruled almost anything in the environment to determine the phenotype of the individual. Now one is interested in what are referred to as complex diseases or quantitative traits that involve both the genotype, of possibly multiple genes, and the environment, which also may interact. These problems are much more difficult. As I said, at each stage when we think we can tackle more ambitious problems, we realize that the number of problems that appear to be solvable has actually grown and not shrunk.

I: What about the total number of genes in the human genome? Is that settled?

Continued on page 11

Continued from page 10

S: I'm skeptical, but the answer commonly given is about 30,000. Only a few years ago, people were guessing 100,000. I would guess 30,000 is closer. But that ignores features that have only recently been recognized as important. One of these is what is referred to as "alternative splicing" so that a single gene, depending on how the pieces of the gene work together, can produce more than one protein. The mantra of molecular biology 50 years ago was "one gene, one protein". Since there are many proteins, one had the idea that there must be many genes. Now it appears that the number of human genes is much fewer but the number of proteins is still very large. So the basic problem doesn't change simply because we now say there are fewer genes. There is still a large number of functions that are incompletely understood.

I: Has any work with the computer led to theoretical insight in your research work?

S: I don't have a very good answer for that. I think that the computer is so much an intrinsic part of my research that it's hard to say what is an insight based on something I've done on the computer or some other kind of insight. It's very easy to say that the computer helps eliminate paths of research leading to dead ends and reinforces fruitful pathways. But working out detailed examples with paper and pencil is the more old-fashioned way to the same result. To some extent, I'm an old-fashioned person. What insights I've had come from piling up individual cases and trying to find the general pattern. I'm very envious of people who seem to get insights without compiling lots of special cases and who seem not to need to do the calculations until they already know what it is they want to calculate. In my case, most calculations are wasted. There's always a pile of papers on my desk. I cover them with scribbles and throw them away very quickly. The computer is helpful in saving some of those efforts in certain cases. Another very important consequence for statistical analysis is that the computer redefines what one means by a solution to a problem. There are still things computers can't do, but basically a problem is solved once it's reduced to something computers can do. Of course, even then, that is not a completely clear answer because what a computer can do in one person's hands is much more than what it can do in my hands. I have the good fortune to work with many good graduate students and younger colleagues, all of whom know computing better than I do. Often they will keep me from spending too much time in blind alleys by doing some computing for me.

I: Is there any software for the application of change-point analysis?

S: People do develop software for change-point analysis. I don't know of any commercial or large-scale programs largely because I don't use such programs on a day to day basis. I'm very poor at using other people's software, so when I want to do some computing I usually write primitive programs of my own. I have seen software that advertises the ability to do change-point analysis but I have never looked at it carefully to decide whether it is the right way or the way I would do it. Software development is a valuable activity, but it's not for someone of my primitive computing skills.

I: How often do you interact with clinicians and medical practitioners?

S: Here we have an issue of the definition of "interact". If interact means to sit down in an office face to face and have an in-depth discussion of a problem, the answer is "not very often", a couple of times a year. If it means to have a more superficial discussion trying to see whether we have common ground for deeper collaboration, then it's certainly much more often. Many of these discussions, I think, don't lead directly to that collaboration, but I find them very useful nevertheless in trying to formulate problems. Often my formulations are fairly theoretical, so I don't try to propose my research as an immediately practical solution; but I find these discussions a very useful conceptual bridge to finding an interesting research problem. If you broaden the definition more to mean reading articles in medical or genetics journals that don't themselves have completely satisfactory solutions to their statistical problems, I would say I spend a great deal of time doing that. That may be one of my primary sources of stimulation in finding problems. When I was much younger, I read the mathematics, statistics and probability literature to improve my techniques in solving problems that were already formulated. Now I depend on other people to tell me if there is an interesting new mathematical or statistical technique, and what I am really more interested in finding out is if there are new scientific problems that are to my taste, which is somewhat idiosyncratic. That may not be what people mean by interaction, but it's interaction at a distance, by the printed page, and I do that a great deal.

I: Do you interact through meetings or conferences?

S: Certainly. Each year I attend a few statistics meetings and a few genetics meetings. The main reason for going to the genetics meetings is to find out the way the science is going and to try to infer what are interesting statistical problems from what people are taking about. These can be problems that they realize they have not solved satisfactorily, or problems where I am not completely satisfied with the proposed solution. In either case I'm often stimulated to try to see what I can do.

I: I think you have touched on a related question: how do you choose the statistical problems you work on?

Continued on page 12

Continued from page 11

S: I have certainly made a transition over the years in the sense that it is now rare that somebody says, "This is a beautiful mathematical problem you will be interested in", and I respond favorably. I'm much more inclined now to respond to the scientific description of a problem that I can see, or somebody will tell me if I don't see immediately, is related to a statistical problem that I might take an interest in. That was probably not the case when I was much younger. Anything that was related to what I was doing mathematically would automatically interest me. Occasionally I still work on problems solely because of their mathematical fascination, but much less so.

I: Has it happened that after attending a lecture or seminar a problem occurred to you and you wanted to solve it?

S: Yes. I don't think very quickly. I'd say, probably two out of three times when I come out of a seminar thinking that I have something to contribute to a problem that was discussed, it turns out I was wrong. Occasionally that can be a useful stimulus to further research. In many other cases, a seminar does not provide a problem that I work on immediately, but gets stored in the back of my mind in case a related idea turns out to be useful. In the world of mathematics people often admit they never understood somebody else's idea until they rediscover it for themselves. I think this is a real phenomenon. You listen to a seminar or hear a series of lectures on a subject without really internalizing it until a few years later when you circle back to this area by who knows what route, view it your own way, reconstruct what somebody was trying to tell you years earlier and for a while even think it's your own idea. Eventually you recognize that somebody else was there first. Maybe you can still make a contribution, or maybe you can't. Of course, one always hopes that one recognizes the situation before trying to publish a paper as one's own idea that was really something learned at a lecture a few years earlier.

I: Do you do direct consultation work?

S: I do a bit, but not much. There are a few people I work with who know the kind of problems I'm interested in and will be good enough not to come to me for routine assistance, but will come with a problem that interests me. This applies particularly to my colleagues at Stanford. Perhaps this is one of the main benefits of having moved there. A fairly large number of my colleagues in the statistics department are involved in many different problems throughout university, and they are kind enough to use me as a secondary consultant by suggesting problems that they know I would be interested in. If the problem originated outside the department, then I will often go directly to the source. This is exciting because the problems are often important, and it's much better for me than working as a real consultant for a living. Then you have to take problems for which there is a flow of income regardless of whether they are interesting or not.

I: Do you get people calling you up to ask whether you could solve this problem for them?

S: Occasionally, but not usually. I have been department chair from time to time and then it happens, not because the person knows anything about me or my interest, but simply because he finds my name somewhere in the directory or on the internet. Then I'm the first layer of contact and I play the role of trying to suggest colleagues who would be most suitable and most inclined to work on the problem insofar as I understand it. That has its own rewards but is quite different.

I: You were Associate Dean of Stanford's School of Humanities and Sciences from 1993 to 1996. What is your most memorable experience in that capacity?

S: I would say that the overall experience was quite memorable, but no single event. My role was to serve as an intermediary between the Dean of the School of Humanities and Sciences (which involves about 30 departments: humanities, social sciences, natural sciences) and the six natural science departments. The reward to me was to learn what was going on in the science departments. Part of the job that I did not particularly like was learning the enormous cost of doing modern laboratory science. I'm very thankful that I am not a laboratory person although I can also see the excitement of doing laboratory work, being closer to the scientific problems than a statistician can be, even for one doing genuine applied statistics. Lab scientists generate lots of data, and without them there wouldn't be any statistical data analysis. But modern science is an enormously expensive business and part of the job of the dean's office is to help allocate resources. You never can make people happy when you are allocating scarce resources. Learning why scientists want the resources and trying to prioritize competing requests is interesting and stimulating. It was fun trying to figure out what different people were doing, where the quality lay, what should be supported or what not. But you can never provide all the resources you want to, and you never learn as much about what is going on as you want to. You sometimes think that if you spend a few more hours, you would really make a better decision. But in the end you are forced by schedules and so forth to make decisions even when you don't understand things completely, and then you can make people upset. There are ups and downs. I'm happy to be back in my role as a scientist, which I find much more interesting.

I: Do you think statisticians are indispensable?

Continued on page 13

Continued from page 12

S: I think they are very fortunate to have the opportunities to play as many roles as they do. They are dispensable if they abdicate their responsibilities to participate in the general scientific enterprise to the extent that scientists find it easier and more satisfactory to do their own statistical analysis. But it probably also works in favor of statisticians that they are very inexpensive. It may not make sense for first rate biomedical scientists to devote a substantial part of their time to thinking about statistics if there are helpful statisticians available. You can have a first grade mathematics and statistics departments with much smaller investment than a first-rate chemistry department.

I: The humanities and the sciences are under the same school at Stanford, but they seem to be incompatible.

S: There is a constant argument as to whether they should be broken up. In the United States, the Stanford arrangement is not unusual, but it is also not universal. One somewhat interesting feature of being an associate dean was to learn about different administrative structures in different universities, and which problems the structures help to solve and which ones they don't solve. For example, I was on a review committee once for the Department of Statistics at the University of Chicago. At that time I was just beginning and spent some time talking to the long-term dean of the School of Physical Sciences at Chicago, which has a quite different structure from our School of Humanities and Sciences. For example, their School of Biological Sciences includes medical faculty. At Stanford there are several "biology" departments, one in the School of Humanities and Sciences and several in the medical school. You would think that certain problems that arise at Stanford might have been solved by the different structure at Chicago. But it seems that while some problems are alleviated others are created, and still others exist with either administrative structure.

I: Were you able to bridge the gap between the scientists and the people in the humanities?

S: For most of my time in the dean's office, my main concentration was on the science departments. I didn't put in as much effort interacting with the humanities departments. For a short time I was put in charge of the philosophy department and the interdisciplinary program on ethics in society. I have occasionally thought that I am a "closet" philospher but fortunate that I don't have to earn my living that way, so I don't have to be rational, or consistent or possess other qualities we expect of philosophers. This was a very interesting experience even though I found it difficult to make informed judgments and came to rely a great deal on telephone conversations with faculty at other universities.