A Journey Through Random Process Data Analysis:

One Type to the Next

David R. Brillinger

$$2\pi \neq 1$$

[Enlarge]
[Look up at audience.]
++++++++++++++++++++++++++++++++

[JOKE]

Introduction.

Today I will be talking about my journey through random process data analysis.

Random processes are fundamental to studying data in time and space and to much work on contemporary scientific problems. They are generalizations of ordinary random variables.

Before focusing on random processes lets talk about statistics generally. Few know what the modern statisticians do. Today's statisticians play key roles enriching the methodology of science - pure and applied. They are concerned with data collection, data analysis, data reduction, data modeling, and data inference. What's more these days nearly everything has become data, and the data are typically stored in a computer. Further the data have gotten steadily more complex. Basic data elements of concern now include graphs, trees, tessellations, shapes, regions, cellular structures, and almost any mathematical object. Statistical dependence is basic and the dimensions of the quantities are much increased. Parameters too have taken on more complex forms, eg. curves or measures instead of simply numbers or vectors. An aspect worth noting is that early results that were put down at the time as excessive abstraction, are now basic to practice. The study of process data and random process models provide a major interface of statistics with mathematics, science and technology.

In the following we will provide examples focusing on: i) data that are segments of curves on a sphere, ii) the efficient combination of several images of the same class of object, and iii) the identification of a system with point process output and time series input.

I am going to start with an example from marine biology.

Example 1. Elephant seals.

This story begins with a telephone call from Brent Stewart of Hubbs Sea World, San Diego. Brent had been given my name, perhaps by Don Ylvisaker, as someone who worked on time series problems. Brent

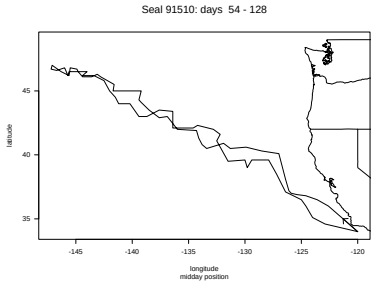

Figure 1: FIGURE 1

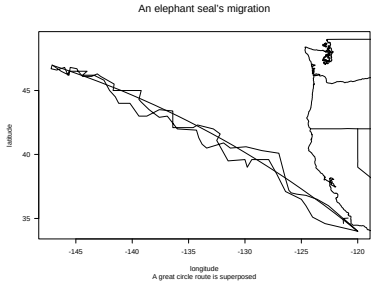

Figure 2: FIGURE 2

was calling with an invitation to attend a workshop on the *Analysis of Time-Depth Recorders* to be held in Fairbanks, Alaska. It turned out that a great deal of data had been gathered concerning the movements of seals of various types, but not much had been analyzed. There were 12 marine biologists and statisticians at the workshop. The topics that the biologists had for the statisticians included: describe migrations, classify dives, identify diving bouts, sampling issues.

At the Workshop I got tied in with Brent to work on the tracks of elephant seals. These seals are exquisite animals. They: were almost extinct a hubndred years ago, they are massive and they dive very deeply. Twice a year they set off from along the California coast on lengthy migrations northwest into the Pacific. Sensors are attached and data for the journeys become available on the animals' return. The figure [FIGURE 1] shows one migration taking 75 days. The points plotted are the estimated midday positions of the animal. These are the basic data for the analysis to be described. As the figure shows the migrations can cover thousands of kilometers.

One of several surprises is that this animal seemed to have in mind both an inbound and an outbound destination. An initial step of analytically describing such migratory trajectories might help to develop some understanding of this circumstance. One possibility is that the seal is approximately following a great circle route, the shortest route between two points on the surface of a sphere. A great circle path has been included in the second figure [FIGURE 2] as a reference. The navigational mechanisms such an animal might employ are as yet unknown, but a great circle route would imply that the animals are able to assess their position relative to some astronomical or global magnetic background and constantly make course corrections. As the elephant seals dive and forage continuously while migrating, there is a need for course corrections.

How might such data be summarized? One approach is to build a stochastic model and thereby be led to pertinent statistics.

Differential equations have been used since the time of Newton to describe the motion of objects. Lately probabilists have been developing stochastic analogs. The case of a particle wandering randomly

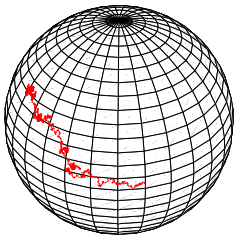

Figure 3: FIGURE 3

on the surface of a sphere was considered by Perrin in 1928.

In the case of migration, a drift term needs to be introduced. For the seals the initial destination may be some point in the sea up near Alaska or perhaps the animals just keep moving ahead. In the case of the return the destination is the starting point on the California coast.

Other aspects of the problem that needed to be taken into account in the analysis were: the positions are available only at noon, there were missing values, and there was measurement error. With some further assumptions a likelihood function could be set down and parameters estimated. A diffusion with drift appeared to fit various of the data sets well. The error of location appears to be more substantial than foraging variability.

Discussion

A random process was made use of here specifically one described by a stochastic differential equation on the surface of a sphere. The data are viewed as a segment of a realization of this process.

The great circle approximation was a surprise and generated various scientific conjectures as to its origin.

In the course of this and related work the importance of *simulation* has shown itself useful. It may be used for program checking, likelihood computation, bootstrapping, and estimation.. A likelihood function was estimated in the present case by simulation. Paths of the fitted process were computed and compared to the ones measured to check the reasonableness of the model. The figure [FIGURE 3] provides an example.

The workshop summary stated

> *"... that the workshop's greatest accomplishment involved connecting marine mammal researchers with statisticians expert in appropriate methods.*"

In the course of the work Brent got a law degree. He became convinced that he needed one to really to help marine life effectively!

The process here was: spatial-temporal, vector-valued. The model involved: an SDE, measurement error. The data analyses involved included: graphs, smoothing, residual analysis.

This example has referred to process data. We turn to a discussion of that.

2. Process data analysis.

The term *process* has historically referred to phenomena which show a continuous change with time, to functions of time. In generalizations *"time"* has become discrete, multi-dimensional, set-valued, and even function-valued. Further "continuity" has taken on extended meanings as well.

Tukey and Wilk have written re data analysis generally,

> *The basic general intent of data analysis is simply stated: to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyzer and recordable for posterity.*

There are a broad variety of *process data* types. Process data refers to information that has been derived by observation of a process at some collection of "time" values. The information will often have numerical form; however, its values can lie in some general structured space.

One list of the *aims of data analysis* includes: discovery of phenomena, modelling, preparation for further inquiry, reaching conclusions in statistical terms, assessment of predictability, and description of variability. These are but a few of the possibilities. We have available today a broad collection of methods for meeting aims. Various factors enter into the choice of method for an intended analysis. One is the degree of urgency involved in the situation at hand. A second is the computing facilities available.

The field of *exploratory process data analysis* consists of the techniques which, when applied to process data, lead to improved knowledge. The purposes include summary, decision, description, and prediction. There are both theoretical and applied sides. Tools employed include: laptops, stochastic models, differential equations and complex algorithms. The tools of data analysis are like those of a mechano set with many components of a broad variety of uses. Special computer data storage methods have been developed.

The aims of data analysis listed above apply to the process data case as well. Having in mind the great variety of process data, we may also mention: control, classification, establishing causation, description of relationship, summarization, removal of concomitant variation, measuring degree of association, signal reconstruction and enhancement, questioning conformity of theory to data, focusing information, precise measurement of constants, comparative analysis.

At the operational level the methods available for process data analysis depend upon the type of process of concern. Any technique employed will depend intimately on the aim of the analysis. Manipulations possible for process data depend upon the computational and instrumental facilities available.

It is clear that one can contemplate working with quadratic and other polynomial forms in the data. This has proved to be successful on many occasions. The great advantage of such forms is that they may be manipulated directly and that computational devices for their evaluation are often available. The step away from polynomial forms is a long one. Experience and insight have sometimes suggested particular statistics to work with. Alternatively, models of the situation of concern have proved a rich source.

Among specific methods applicable to process data are: spectrum analysis, smoothing, inversion, likelihood, Kalman-Bucy, clustering, re-expression, dimensional reduction, contingency, analysis of variance, least squares, simulation. Specific algorithms exist for their application to many types of data.

Problems continue to be the classical ones of: summarization, model assessment, display, including explanatories, efficiency, robust variants, sanctioning, and missing values. New ones arising include: How to get experimental output of special form into a computer for analysis? How to display and make available unusual data types? In the search for insight and unusual occurrences (eg. outliers), descriptive statistics have returned as a force.

Example 2. Electron microscopy.

As a second example we turn to data that are images of a type of biological object.

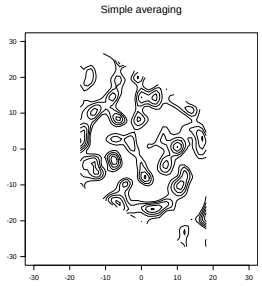

Figure 4: FIGURE 4

The story begins with my biking home one day in 1987. I caught up with a neighbor Bob Glaeser. The two of us began walking along together. We were aquainted from having sons in Scouts. At one point Bob asked,

"Do you know anything about the Fourier transform?"

I muttered a response and a collaboration experience began.

Bob is a biophysicist, specializing in structural crystalography. Amongst other things he is interested in efficiently combining individual images of crystaline substances to obtain improved images. He obtains the images via an electron microscope.

Electron imaging is a technique that biophysicists employ to learn about molecular structure, particularly the placements of atoms within molecules. The work is carried out mainly with materials that are crystalline. Then one has regular replications of the same basic entity and one can enjoy the benefits of averaging.

The substance studied was a protein, *bacteriorhodopsin*. The units are tiny, measured in Angstroms. Photographic films were digitized by a scanning densitometer. The figure [FIGURE 4] shows an estimated unit obtained by averaging about 160 units. This computation was carried out by Fourier analysis taking advantage of the periodicity.

The problems, brought to me by Glaeser, were to improve such images by statistical methods, to efficiently combine several images of the same material, and lastly to assess uncertainty in the final estimates. Improvements would result from taking note of the presence of noise in an image and of the noise's distributional character.

The structure of a crystal being periodic, it was convenient to construct a representation of a unit by a Fourier series expansion. The researchers knew that they could obtain good estimates of the amplitudes of the coefficients via diffraction experiments. Fourier amplitudes were therefore extracted from electron diffraction patterns. The earlier experiment was one in imaging. These amplitudes could be inserted in the Fourier expansion replacing the amplitudes obtained from the imaging experiment. There was yet another trick that the biophysicists had developed. Blow and Crick were lead, by a Bayesian argument, to insert multipliers times the Fourier coefficients. So besides the averaging, multipliers were employed and found to further improve the images. The figure [FIGURE 5] shows an image obtained in this fashion. It shows considerably more detail than the previous image.

There remained Bob's questions of how to combine a number of images and how to indicate the estimate's uncertainty. To address this it was noted that the discrete Fourier transforms of discrete planar

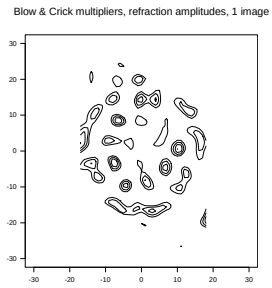

Figure 5: FIGURE 5

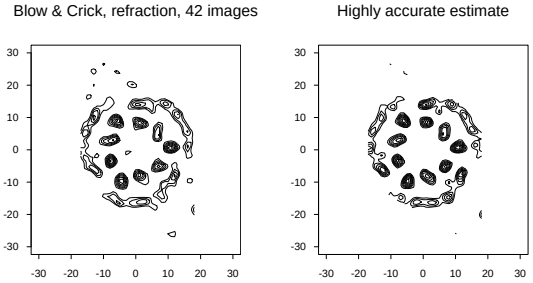

Figure 6: FIGURE 6

fields are sums and thereby central limit theorems suggest approximating the distributions by the normal. Normal theory can be invoked to justify weighting the coefficients from different images inversely proportional to their estimated variances. At the same time statistical subject matter could be invoked to recognize the Blow-Crick multipliers as being shrinkers and thereby alternate improved forms could be considered.

The next figure [FIGURE 6] shows the result of combining 42 images. A very substantial structure appears. For example one notes the 3-fold symmetry. The procedure may be further validated by comparing the result with a highly accurate image obtained by other means. This is the righthand image in the figure. There is impressive agreement.

There remains the issue of how to compute and display a measure of uncertainty. The next figure [FIGURE 7] provides 10 simulations of the image superposed. The simulations have been generated employing the Gaussian model that was fit. This procedure was particularly convenient given that the images were contour plots.

Discussion

To quote Bob Glaeser,

> *"... it can be anticipated that this merging scheme will be extremely valuable for many types of biological 'problem' specimens, ..."*

Data analyses that were carried out include: Fourier inference, simulation, ...

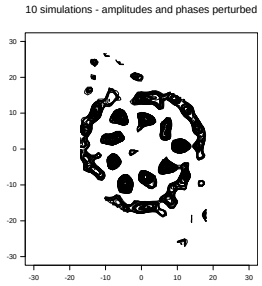

Figure 7: FIGURE 7

Bob has recently won the ? Prize of the American Society of Crystallographers.

Fourier analysis has been referred to. It is a powerful tool for process data analysis. By Fourier transforming is meant using sines and cosines to isolate frequency-side effects, at least approximately.

John Tukey once remarked,

"One can FT anything, often meaningfully."

I sat in a Princeton classroom in 1963 and watched John Tukey come out with a Fast Fourier Transform (FFT) algorithm. I even reviewed the notes! Missed an opportunity there. Later Gordon Sande and J. Cooley seized the problem and prepared Fortran subroutines.

Concerning a particular Fourier inference method Tukey has written,

> *"My work in spectrum analysis - ... - has been an active catalyst for my ideas about data analysis in much more general contexts. A good reaason for this is spectrum analysis's having all the troubles of other data analysis, only more so."*

There are other transforms that also have fast computing algorithms and are proving useful these days. An example is wavelets.

The example contained random processes. It is time to talk about them.

Random processes and statistical methods.

The statistician typically proceeds by building a *chance model* for a situation. The vast majority of statistical analyses rest on a *probability*. Consideration of a random entity allows all of *probability theory* to be brought to bear on problems in particular, for example, results concerning special random processes. An ubiquitous concept in the work of statisticians (and indeed of all researchers) is that of a (stochastic) *model*. A variety of meanings are attached to the word. It is often taken to mean a theory. With a model at hand, much of a researcher's work becomes deductive and manipulative. The greatest difficulties lie in creating pertinent models. Statisticians end up with a schizophrenic attitude to them.

Martin Wilk has said that

> *"The hallmark of good science is that it uses models and 'theory' but never believes them."*

We are now in the era of random process data analysis. Just what is a random process? It is simply a family of random variables or chance quantities. However today's usage often seems to have in mind

that the index labeling the family has something extra, like an order or a toplogy or an algebraic structure. Often the index refers to time or space or both.

The idea of a random process need not be complicated,

Simply pick a realization out of a hat full of them.

Alternately a random process is a probability measure on a function space. In operational use a random process is a random function, or random measure, or random generalized function with domain that is temporal or spatial or spatial-temporal. Its values have coordinates. Its realizations are: curves, surfaces, shapes, figures, sequences and the like. It relates to situations where things move and change - things in time or space or both.

Today's move to the empirical analysis of random processes is made possible in part through the development of useful probability models for unusual mathematical structures and the availability of new devices and ideas for the storage and manipulation of random process data sets.

In developing inferential approaches a *random process datum* is often assumed a part of a realization of a random process.

There are various types of random processes including: *White noise, Markov chain, Stochastic point process, Branching process, Birth and death, Gibbs process, Diffusion, Stochastic geometry - line process, fibre process, random set, tessellation*

There are also *random process systems*. These are structures consisting of inputs, an operation and corresponding outputs. There now exists an immense literature concerning ısystemt identification given data consisting of pieces of (process) input and corresponding pieces of (process) output. This is the random process analog of the regression analysis of ordinary statistics. There are random process variants. An essential practical distinction arises between situations in which the scientist can select (some, of) the inputs and those where they are outside his control. Another distinction is whether the model is *mechanistic/conceptual* (based on specific description of the natural components involved) or empirical (based on regularities that caught the researchers eye). The former is the fundamental one.

One of the major contemporary works on statistical inference for random processes is that of U. Grenander, *Abstract Inference*. It is worth indicating some of the distinctions he recognizes and problems and procedures that he highlights. By his choice of the term "abstract inference" he deliberately leaves ambiguous whether he means the sample space (set of possible observations) or parameter space (values for quantities characterizing the probability distribution at hand) or both to be "abstract". In the work he discusses each case. For inference he employs: linear methods, likelihood based estimates and direct methods (the latter being based on common sense estimates). Related circles of ideas include: penalized maximum likelihood, Courant regularization, Bayesian estimation, ridge regression, and Stein estimates.

Because of the difficulty of the problems being studied and the massive amount of subject matter, there is a basic need for collaboration with scientists from other substantive fields.

Some of the key scientists contributing broadly to random process theory and applications include: Cramer, Grenander, Tukey, Bartlett, Rosenblatt, Parzen, Donoho, Kolmogorov, Hannan, Wiener, Lewis, D. G. Kendall, Akaike

Example 3. Neuron firing.

The final example concerns the identification of a nerve cell system. Pepe Segundo, a professor in the Brain Research Institute at UCLA, came up to talk in a seminar on point processes being run by Peter Lewis and David Vere-Jones in the early seventies. In the course of his presentation Pepe indicated that he had many data sets of neurons' firing. I asked if I could work with some of that data to try out some

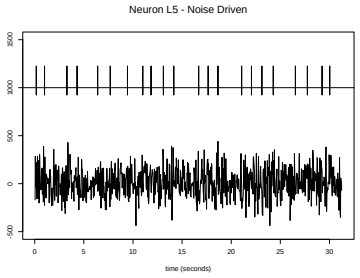

Figure 8: FIGURE 8

methods that I had been developing for point process system identification. A steady stream of boxes of punched cards began to arrive at Berkeley.

In the experiment to be focused on just now a neuron, L5, of *Aplysia californica* was stimulated to fire repeatedly by a continually fluctuating current. The datum collected was a segment of a hybrid process, a time series and a contemporaneous point process, see the figure [FIGURE 8].

The methods that I had been working with sought to assess specific quantitative relationships between the input and output. In essence the models were point process analogs of correlation and regression models.

That work led to a success story namely we were able to infer the ”wiring diagram” of networks of three neurons via partial coherence analysis.

The particular work to be discussed just now though is the building of a conceptual model for a neuron’s firing. The model seeks a probability/intensity function for the firing of the neuron as it depends upon the input. Consider the following model. An electrical input contributes, by summation, to the membrane potential within the neuron. The neuron fires when the potential crosses a threshold at the trigger zone.

There are other neural phenomena that need to be included, for example the refractory period, where a neuron once it has fired refrains from firing again for a time period. This is handled by after a firing, setting the threshold to a high level and then letting it to die off.

The character of the summation function affects whether the neuron is excited or inhibited by the stimulation.

For computations time was discretized to integer time values. The point process then became a $0-1$ time series.

Assuming the threshold to be gaussian a likelihood function can be set down and employed to estimate the summation function and other quantities. The results are presented in the figures [FIGURE 9, 10, and 11]. Figure 9 presents an estimated linear kernel. Figure 10 presents an estimated quadratic kernel. These two together add to give the membrane potential. Figure 11, bottom, provides an estimate of the membrane potential while the top graphs the actual firing times. One can note large values of the estimated membrane potential near various of the firing times.

Discussion.

System case/graphical models

The parameters have conceptual interpretations

The model allows spontaneous firing

Data analyses employed included: ...

The model can have a term to handle nonstationarity/trend

Validation was carried out by ...

Pepe has remarked, concerning such statistical work, that

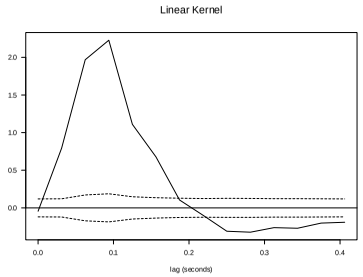

Figure 9: FIGURE 9

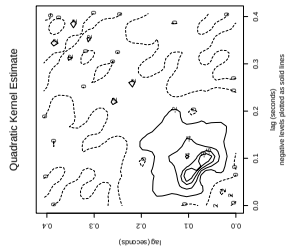

Figure 10: FIGURE 10

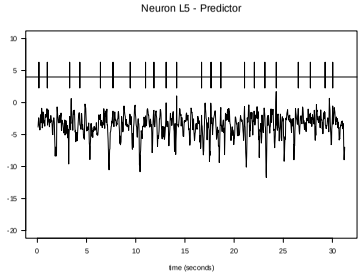

Figure 11: FIGURE 11

"... the ultimate goal is understanding in strictly biological terms."

Difficulties arising.

However, there are continual difficulties that arise in practice and complicate the use of the algorithms. These include: missing data, out-of-line data values, measurement error, concomitant variation, extra structure in the data, artifacts, heterogeneous data, censored data, biased collection procedure, jitter, discretization error, missing values, unequal spacing, long-range dependence. A broad variety of procedures now exist for dealing with these difficulties. There are problems of interpretation caused by urking variables, and nonidentifiability.

Summary / Conclusions / Future.

Return to opening - The talk has focused on process data analysis and presented three applications.

There needs to be combinations of both physical and statistical reasoning.

Subject matter plays essential roles in the analyses made in both the interpretations made and conclusions drawn.

The future

Searching for music, photos, ... on the web (search engines)

Data mining

What was that all about?

The universality of statistical methods/techniques.

The same methods play central roles in the analysis of data from disparate fields.

Statistics contributes efficiency, uncertainty measures

Interactions with scientists

All of statistics seems to be part of any real statistical problem.

Lewis Mumford on abstraction

Abstraction in many fields

There have been stages and stages of generalization

Bell Labs realization that sound, images same

Bring back datum as the singular - n = 1 in r.p. analysis

Advice people have given me through the years:

Martin Wilk *"Go back to the beginning"*,

Neyman *"Say what you are going to say, ..."*,

Tukey *"Figure out at what time of day you are most productive."*,

Kjell Doksum *"Always take a default."*.

Bill Williams *"Buy apartment buildings."*

?. Acknowledgements

JWT, scientific collaborators, Pepe, Brent, the guys at Bell Labs, Bruce, Hilgard, Glaeser, Forest Service

It seemed that everyone I worked with brought something special into my life