The Use of Mathematics for Deciphering the Movement of an Object: A Historical Review of the Introduction of AIC

Hirotugu Akaike

Introduction

The concept of an information criterion was formulated not through studies that followed traditional, classical statistics, but through persistent efforts to get to grips with practical problem solving, and thus it appears to be an unconventional idea. I would like to look back on why I took such an approach in relation to my way of viewing things and the influence of the environment and times I grew up in. I will also chronologically review the introduction of the AIC and its impacts.

Personal history

I was born in a rural area on the southern side of the foot of Mt. Fuji. One of my uncles, who was a pilot in the early days of commercial flight, gave me steam engine and motorboat models. As a child, I loved to turn them upside down to see how they worked. Since then, I guess, rather than observing an accomplished result, I have been more interested in exploring why such a result is produced. I like to try out different things predicting the outcome. Even today, I still maintain this habit in anything I do. "As the twig is bent, so grows the tree."

As an elementary school pupil, I was good at solving questions of applied arithmetic but not good at memorizing or rote learning, such as writing kanji characters, or calculating with an abacus. Influenced by my brother, who lived with the pilot uncle in order to go to school, and another uncle who was a naval warplane pilot, I had an interest in airplanes by the time I went to junior high school. This naval pilot told me about mathematics, in which I became interested.

My English teacher at junior high school was very enthusiastic, but I was not

good at memorizing English words. Instead, I memorized simple sentences, which showed how English words were actually used. I had a passion for kendo and swimming as well, and eventually I entered what was then the Naval Academy.

At the Naval Academy, I had opportunities to learn introductory science and engineering, statistics, and probability. I believe that I learned the usefulness of applying all pieces of knowledge organically through experience.

Aiming to contribute to postwar reconstruction, I went home after World War II. However, I was greatly distressed to see the collapse of social values. After thinking many things, I realized that respecting my own and others' lives was the basis of morality, and this restored my inner peace. Although I attended the Science Course of the former First Higher School, I was concerned with my future career. Eventually, I went to the Department of Mathematics, in the Faculty of Science at the University of Tokyo. This period was a turbulent time for me.

In 1952, I joined the Institute of Statistical Mathematics of the Ministry of Education, where I aspired to work on studies focusing on real issues concerning Japan. After ten years of exploring solutions to specific problems, I had the opportunity to deal with an issue that was directly associated with production activities that supported postwar reconstruction efforts. Since then, I have always been dedicated to problems associated with analysis and control of phenomena that vary with time. I believe this was a result of the interest in the mechanisms of moving objects I'd had since childhood. The choice of this issue required a variety of ideas, and resulted in the creation of the AIC.

Let me move on to AIC-related topics.

Mathematics of prediction

Probability

The effective use of observational data using the AIC requires a model that explains how such data can be obtained. It is usually impossible to precisely describe the

mechanism for generating data, and therefore, we create a model using probability to express this uncertainty.

The concept of probability is quite natural; everyone uses it almost unconsciously for processing uncertain things. Probability uses numbers to express the likelihood of a certain outcome occurring; for example, we say the probability of rain tomorrow is 0.7, or 70%. In this case, the odds are ten to seven that it will rain tomorrow. This seems very simple, but we must think of many factors to estimate the probability of an outcome in reality.

In contrast, the mechanism to determine probabilities is objectively given in some cases, such as the probability of getting a six when you throw a dice. This type of probability is given using a randomizer——a tool that generates random numbers used in fortune-telling and gambling——and from which the idea of probability originates.

In betting on dice, where the prize money depends the number on each side, the expected value of prize money is obtained by multiplying the amount for each number by the probability of getting each number, and totaling them. The key point here is that the expected value is based on probability. Therefore, the participants in the betting think carefully about the probabilities of winning.

The relationship between probability and statistics

When you use probability for the effective development of a future plan, you will find out how complicated probability is. When someone says that the probability of rain tomorrow is 0.7, the value is based on some evidence. For example, the evidence may be the proportion of the days on past record when the weather was like today's weather and it rained on the following day.

Looking at past records, the ratio of rain on the following day was 70 days out of 100 days, when the weather was similar to today's weather. In this case, in ordinary terms, based on statistical records from the past the ratio of rainy days to all days is 70 to 100, or 0.7. In other words, the probability of rain tomorrow is given as an average

incidence.

In short, the mean value of statistical records in the past equals the expected value in the future. This represents the basic, practical idea of probability, which suggests that probability is naturally associated with the use of statistical data.

The basis of this idea is that probabilities can be derived by recording data infinitely, like throwing a dice many times. However, this is impossible in reality. Instead, we use data at hand to create a mechanism of probability that generates future observational values, namely, a model. We make predictions using this model.

In this way, creating a mechanism to forecast future values using available data is the essence of statistical data processing. It is the materialization of what Confucius said: "Study the past if you would divine the future."

Let me continue to speak about examples of the practical applications of probability.

Application to real problems

Statistical control of the silk reeling process

The first example is statistical control of the silk reeling process, achieved in cooperation with Mr. Akinori Shimazaki of the sericulture laboratory of the then Ministry of Agriculture and Forestry. Raw silk yarn was one of Japan's major exports before World War II, and sericulture is a traditional industry with a long history. Production of raw silk yarn largely depended on experience. After World War II, however, statistical process control methods were introduced, triggering the introduction of control chart methods that helped detect abnormalities in processes based on ever-changing observational values.

In those days, raw silk yarn was made by intertwining fibers from a specific number of cocoons. What was controlled was the number of times that fibers were broken and cocoons dropped within a specific time range. When the value increased abnormally, the process was judged as defective. Ordinary control chart methods use

operating records of a machine in a steady state to determine the mean value of the dropping of cocoons and a permissible range of fluctuation, judging the process as defective when the observed value goes beyond the permissible range.

However, data for the statistical distribution of length of a fiber are obtained as a result of experimental silk reeling to identify how to boil cocoons so as to facilitate reeling fibers off the cocoons. Using these data, the likelihood of a fiber breaking in the process of reeling is stochastically determined. This procedure verifies the characteristics of the ideal operation of a machine and theoretically determines the values necessary for control of the process [Fig. 1]. In this way, we achieved remarkable success in practical process control. A piece of past knowledge obtained through experimental reeling was effectively used by thinking about the stochastic structure of a future event, which was the dropping of cocoons.

Estimation of a power spectrum

This is a kind of statistical analysis of random vibration, which attracted considerable attention in the 1960s, and a practical application of a method to measure the characteristics of irregular fluctuation. I worked on putting this method into practice in cooperation with Mr. Ichiro Kaneshige, then of Isuzu Motors, Limited.

When time-series data varying with time are broken down into periodic vibration components, a sharply fluctuating zigzag line that expresses the strength of components along the x-axis indicating frequency (the number of vibrations within a unit time) is obtained. This is called a periodgram, which has an extremely conspicuous component in some cases. The periodgram has long been used for detection of periodic components of geophysical, observational values.

This analysis, however, provides only a zigzag when using data of irregular variation such as vibration of an automobile. Therefore, it was doubted whether this analysis method could be applied to the statistics of the time. Local averaging of data along the frequency axis, however, manifests a power spectrum that presents a smooth fluctuation pattern. As is the case with a light spectrum, a power spectrum indicates

which frequency has a stronger periodic component [Fig. 2].

Estimation of a frequency response function

A further advanced application of probability provides a method to measure the characteristics of a system, such as measuring the steerability of an automobile by analyzing the random motion in a steering wheel. A frequency-response function represents the characteristics of a system in this case [Fig. 3].

This measurement method enables easy measurement of the characteristics of a system in a state that is similar to real movement. This measurement method was successfully put into practice in cooperation with Mr. Yasufumi Yamanouchi, the Transportation Research Institute of the then Ministry of Transport. This estimation method of a frequency-response function was one of the most advanced achievements in the world and was applied to actual cases by researchers in many fields. Results of such applications were reported in an English journal of the Institute of Statistical Mathematics in 1964. Those cases include applications to automobiles, ships, railway vehicles, aircrafts, piping systems, and water-piping systems of hydroelectric power plants.

Optimal control of production processes

I tried to apply the estimation method of a frequency-response function to the automatic operation of cement kilns in cooperation with Mr. Toichiro Nakagawa, at the then Chichibu Cement Co., but a difficult problem came up. Instead, we decided to use a new optimal control theory that is suitable for computer control [Fig. 4, 5].

Professor Rudolf Emil Kalman's system theory——the Professor was a laureate of the 1985 Kyoto Prize in Advanced Technology——served as the foundation for design of the control system. However, we had various difficulties putting Professor Kalman's system theory, which used stochastic expressions derived from the study results of Dr. Kiyosi Itô, a laureate of the 1998 Kyoto Prize in Basic Sciences, into practice in a real, largely variable process. To overcome such difficulties, we began by

formulating a prediction equation for the movement of a process and then thought about how to control a process using this equation.

Autoregressive model

The autoregressive model predicts a future value by adding a current value to a sum of past values multiplied by specific coefficients in a time series of all related variables. I decided to use the following method: I estimated the structure of the autoregressive model using time-series data, and then using the results I analyzed relationships among variables. Prediction errors in this case have the structure expressed as a signal of a simple randomizer called white noise, which enables simulation of the movement of an actual system. Use of white noise makes it possible to predict future, momentary movements of a system to determine appropriate inputs for controlling the system [Fig. 6].

I had a problem here again. It was how to determine the order of a model; in other words, how far you go back to the past to incorporate data into the prediction equation. No practical solution to this problem was found in those days. When the order is too low, the equation is less likely to predict a future. When it is too high, insufficient data causes inaccurate prediction. Therefore, I created a method to compare and evaluate models with different orders. This method enabled choosing the optimal model and, thus, designing control systems.

Making optimal control happen

The multivariable autoregressive model made it easier to grasp the movement of actual systems, helping put optimal control theory into practice in control of cement kilns with irregular variations.

As a result of publishing the calculation program necessary for practical application of this method, verification of the structure using observed data and examination by simulation were made possible and produced outstanding results in various areas where verification of the relationship of movements among variables had

been difficult, such as temperature control of boilers at thermal power plants, automated steering of ships, analysis of a living organism's function for maintaining its internal environment, or homeostasis, analysis of brain waves, time-series economic analysis, and analysis of noises of nuclear reactors.

The application to steam temperature control of thermal power plants was made possible by Mr. Hideo Nakamura, then of the Kyushu Electric Power Company, Inc., with help from researchers at the Central Research Institute of Electric Power Industry [Fig. 7, 8]. When I had an opportunity to observe the restart of operation of a boiler after a regular inspection, the chief operator of the boiler welcomed the introduction of the control method. I also heard from Mr. Nakamura later that the master of control theory, Dr. Yasundo Takahashi (1912-1996), had thought highly of this control method, saying that the sun rises in the east and optimal control of actual processes comes from Japan.

Since then, this control system has been used by Japanese and overseas boiler manufacturers and is in operation in China and Canada as well as in Japan. This is a valuable, successful case that connected theory to reality, achieving application of optimal control theory to actual processes with wide variations, as well as the application of the automatic operation of cement kilns.

To tell the truth, it was not so easy to determine what to use as the evaluative value in choosing the order of the basic model. During the process of solving this problem, I was attracted to applying the idea of likelihood. I then came up with the information criterion, or AIC, of which I am going to speak now.

Clarification of likelihood

What is likelihood?

I believe many of you are not familiar with this word, so I would like to provide a brief explanation of likelihood that evaluates stochastic quality of models. This is a very interesting idea, based on which an information-quantity criterion is defined.

Suppose that you throw a dice and get a six. A question then arises [Fig. 9]. There are two dice, dice A and dice B. It is certain that one of them was thrown, but it is not known which was actually thrown. In such a case, you must judge which dice was thrown (A or B). This is like a case where a private detective is trying to identify a criminal between suspects A and B using an observational value, or six.

Suppose that the probabilities of getting a six when throwing dice A and dice B are P(6/A) and P(6/B), respectively. If P(6/A) is much larger than P(6/B), it is natural to think dice A was thrown when the six came up. When generalizing this idea, P(x/A) and P(x/B) for observational datum "x" are called the likelihood of dice A and dice B for datum x.

In this case, we must note that datum x has already been observed and is definite. P(x/A) and P(x/B) are the likelihood of dice A and dice B showing a six, respectively, but they are not probabilities of getting a six.

Probability uses past knowledge and experience to predict future data, while likelihood uses current data to evaluate a framework that generated such data in the past. It seems that we are discussing how to handle knowledge beyond Confucius' words: "Study the past if you would divine the future." The essence of this question will be clarified in the next discussion on information quantity.

Information—quantity criterion

Information quantity

One of the rationales for using likelihood as the value to evaluate a model is the use of a quantity called information quantity. Information quantity I(Q : P) measures the distance from model P to real structure Q.

Here P and Q stand for stochastic mechanisms that determine how observational values are generated, and these mechanisms provide probabilities of getting datum x as P(x) and Q(x), respectively.

When datum x has been observed, the natural logarithm log P(x) is used to

determine the value of I(P : Q), where P(x) is the likelihood of P. This is called log likelihood. Even though we do not know the true structure Q, this conversion explains from the perspective of probability that it is reasonable in the comparison of two models to judge that the one with a higher log likelihood is more likely to be Q. A model with higher log likelihood is deemed to be closer to the true one, and an analogy for this is a mountain with a higher altitude being closer to the sky [Fig. 10]. This is the meaning of an information-quantity criterion, which clarified the rationale for the use of likelihood.

Models that include parameters

In an actual situation where we develop inferences using likelihood, we use models that include some adjustable variables, or parameters. When parameters are adjusted, the structure of a model changes. Consequently, the probability of getting the same conclusion as the current observed value, in short, the likelihood, can be adjusted.

A simple and concrete image of this is given by a dice with a heavy iron ball in its barycenter. Suppose that the position of the ball can be adjusted back and forth, from side to side, and up and down using three adjustment screws, or parameters. If you move these screws, it will increase the probability of the dice landing on the side that is closer to the iron ball. In this way, you can obtain a model that can adjust probability distribution [Fig. 11].

AIC

If you use a model with some parameters and apply the maximum likelihood method in which parameters are adjusted to maximize the likelihood for a given observational value, the model can be assessed using an information-quantity criterion, as shown in this equation:

AIC = (-2) (the maximum log likelihood) + 2 (the number of parameters)

A subtle feature of the AIC is to use the log likelihood instead of the likelihood itself.

The larger the AIC is, the worse the model is deemed, due to the coefficient of

-2 to the maximum log likelihood. The second term, or "2 (the number of parameters)," is obtained by adding the corrective value for an increment of the likelihood due to adjustment of parameters to the evaluation of errors in the prediction using the applied model. This term prevents an increase in unnecessary parameters.

Publication of AIC

The AIC was first reported at the meeting of the Japan Statistical Society in 1971, and later in the same year, I presented its details at the 2nd International Symposium on Information Theory held in the Republic of Armenia. The presentation was published in 1973. In 1974, I contributed a review of the article to the bulletin of the Control Systems Society of the Institute of Electrical and Electronics Engineers, Inc., U.S.A.

There was positive response from the applied fields, but muted reaction from the classical statistics field. However, in 1992, the 1973 article was published in a book that put together some remarkable, theoretical contributions to contemporary statistics. This shows that the AIC had gradually been accepted.

At present, philosophers are also interested in the AIC. In logic, there is a principle called "Occam's razor," that does not use unnecessary assumptions. The term "2 (the number of parameters)" of the AIC prevents an increase in unnecessary parameters. In this context, some people have developed an argument considering the AIC to be the materialization of this principle.

However, the precondition for the use of the AIC is to propose a model. Modeling requires effort to repeat hypothesizing and hypothesis testing infinitely using objective knowledge, empirical knowledge, and observational data. C. S. Peirce proposed the idea of inference aimed at achieving truth through such a hypothesizing process and emphasized this type of inference, naming it abduction, or retroduction. The AIC helps carry out this process.

The effect of the introduction of an information−quantity criterion

The theory of the maximum likelihood method has been discussed as a way of using observational data to forecast values of parameters in the true structure. But in practice, the maximum likelihood method can only be applied to artificial, experimental cases. For practical use, the maximum likelihood method is applied to an artificial structure that is aimed at extracting useful information from data; in other words, a model. Thinking in these terms, we may realize that elimination of subjectivity, as emphasized by classical statistics, in fact restricts the flexible thinking needed to develop and use a model, and as a result prevents the realization of a useful data-processing method.

The introduction of an information criterion brought about flexible thinking, which increased the degree of freedom in creating a model. Consequently, usable models including those based on psychological images rapidly increased in kind. This effect can be confirmed in the example of the seasonal adjustment method that uses Bayes model to separate impacts of seasonally varying components from an economic time series [Fig. 12]. This figure illustrates the concept of Bayes model [Fig. 13].

This model has also been used in the development of useful processing of geophysical observation data obtained from analysis of earth tides and other natural phenomena. Let me show you an example of outputs from a calculation program compiled by Mr. Makio Ishiguro from the Institute of Statistical Mathematics, and Mr. Tadahiro Sato and others from the Mizusawa VERA Observatory of National Astronomical Observatory of Japan [Fig. 14].

Conclusion

In closing, let me talk about effective use of knowledge. There is a well-known story with respect to the practical use of the airplane. Scientists used their scientific knowledge to prove the airplane could not be put into practical use. In contrast, the Wright brothers used their knowledge to fly in the sky and put the airplane into practical use. This is the case with the AIC, too. I realized the limitations of data processing based on the existing theory. I examined the existing theory, according to the requirements of problems, and then achieved the AIC.

You might have an impression that my story about the AIC started with how I had a connection to the airplane in my childhood, and ended as a comparison of studies on the AIC with studies on the airplane. However, in closing, what I want to emphasize is the importance of having a sense of purpose, or trying to achieve something. In achieving control of cement kilns and boilers at thermal power plants, the researchers were remarkably persistent and energetic. Hoping for deeper understanding and interest in this regard, I would like to end my talk.

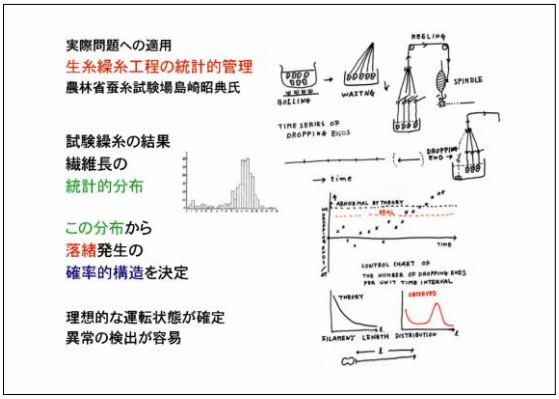

Fig. 1

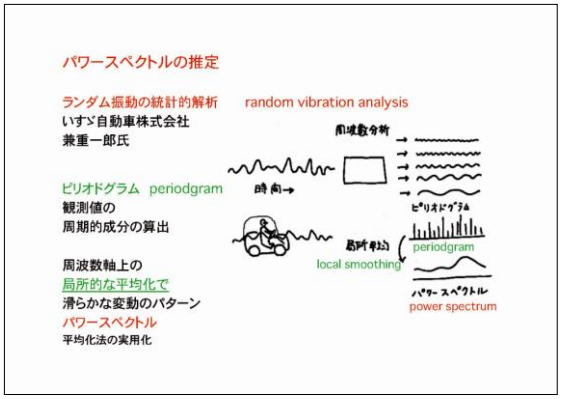

Fig. 2

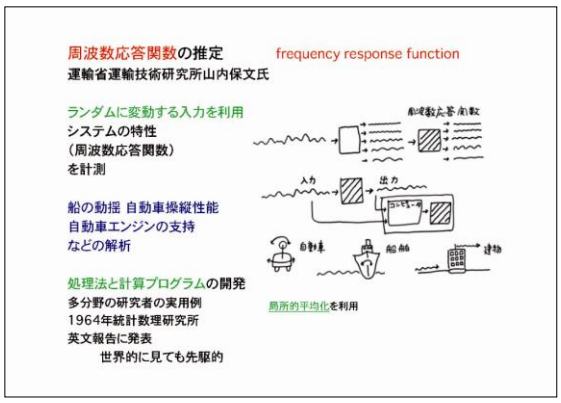

Fig. 3

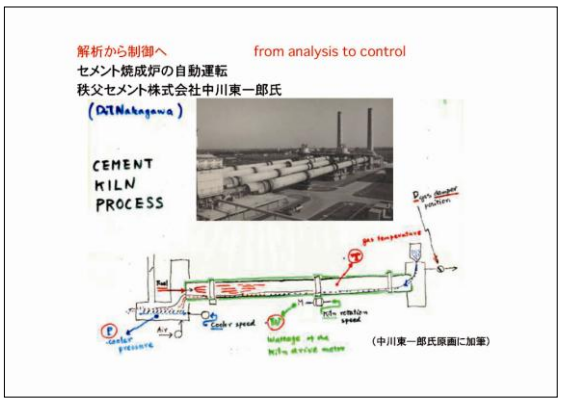

Fig. 4

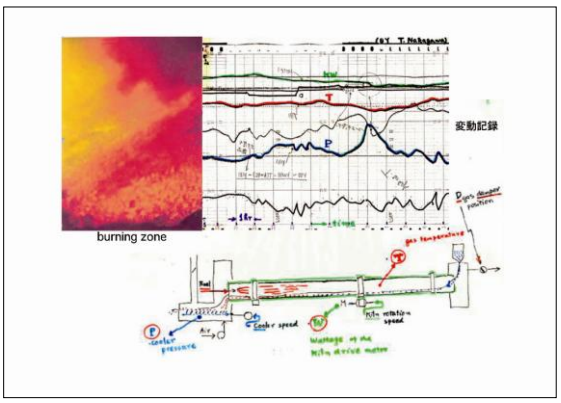

Fig. 5

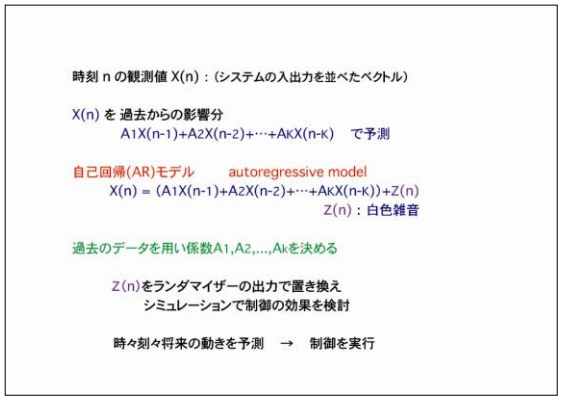

Fig. 6

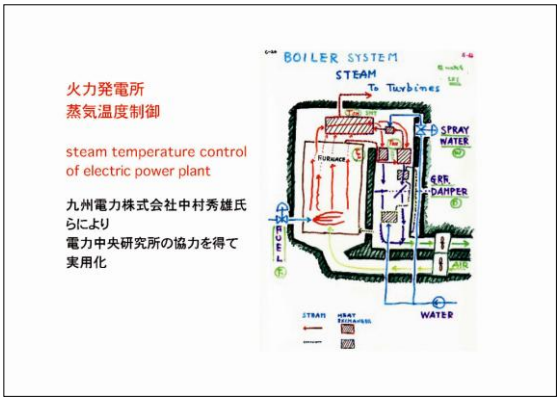

Fig. 7

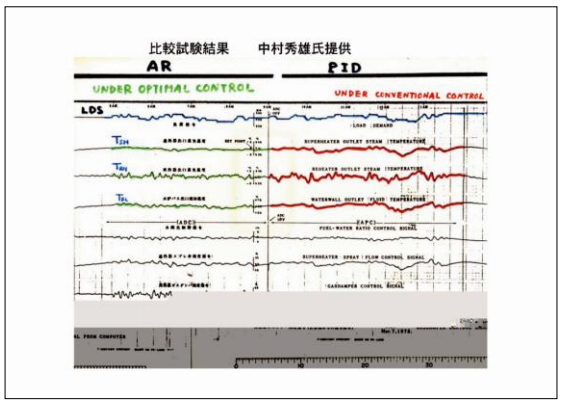

Fig. 8

Fig. 9

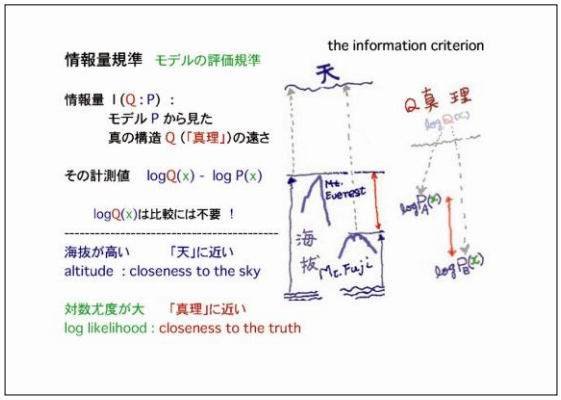

Fig. 10

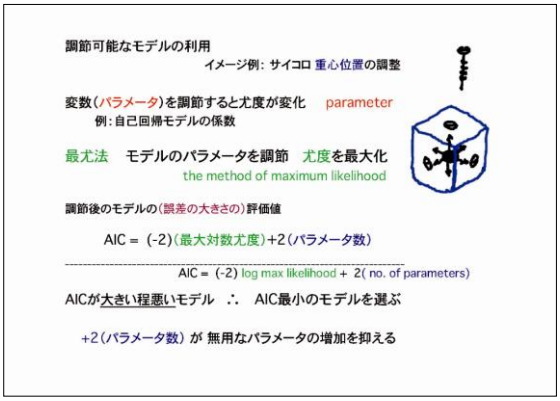

Fig. 11

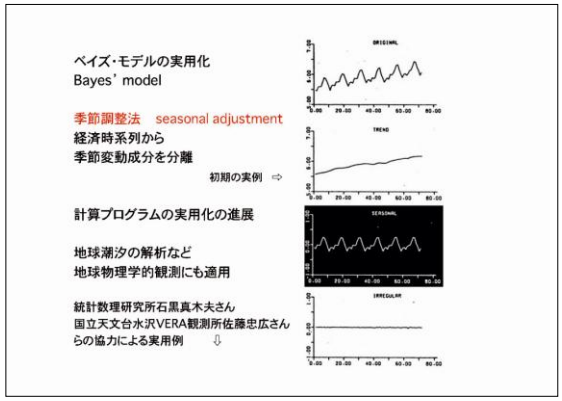

Fig. 12

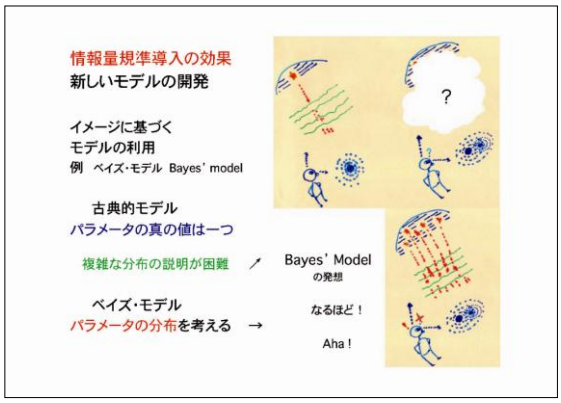
情報量規準導入の効果
新しいモデルの開発
イメージに基づく
モデルの利用
例 ベイズ・モデル Bayes' model
古典的モデル
パラメータの真の値は一つ
複雑な分布の説明が困難
ベイズ・モデル
パラメータの分布を考える
Bayes' Model
の発想
なるほど！
Aha !
?

Fig. 13

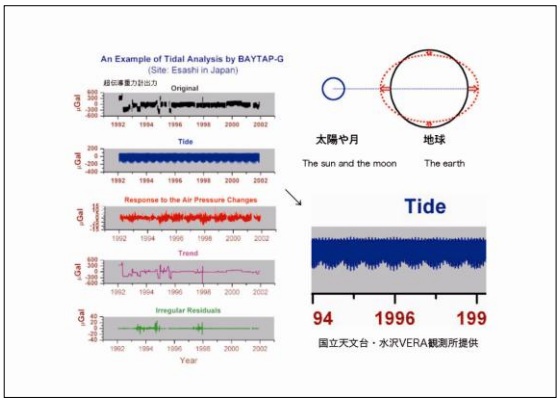
An Example of Tidal Analysis by BAYTAP-G
(Site: Esashi in Japan)
Original
Tide
Response to the Air Pressure Changes
Trend
Irregular Residuals
Year
太陽や月
地球
The sun and the moon
The earth
Tide
94
1996
199
国立天文台・水沢VERA観測所提供

Fig. 14