Prospective Scientific Methodology in Knowledge Society

Genshiro Kitagawa

The Institute of Statisical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 Japan
kitagawa@ism.ac.jp

Abstract. Due to rapid development of information and communication technologies, the methodology of scientific research and the society itself is changing. The present grand challenge is the development of the fourth methodology for scientific researches to create knowledge based on large scale massive data. To realize this, it is necessary to develop a method of integrating various types of information and of personalization, and the Bayes modeling is becoming the key technology. In the latter half of the paper, several time series examples are presented to show the importance of careful modeling that can take into account of essential information.

Keywords and Phrases: Information society, knowledge society, data centric science, active modeling, time series analysis.

1 Change of Scientific Research and Society Due to the Development of Information Technology

1.1 Change of Society Due to Informationization

Due to the progress of information and communication technologies (IT), large-scale massive data are accumulating in various fields of scientific researches and in society. As examples, we may consider the microarray data in life science, POS data in marketing, high-frequency data in finance, all-sky CCD image in astronomy, and various data obtained in environmental science, earth science, etc.

Rapid development of information and communication technologies influenced the research methodologies of science and technology and also society itself. In the information society, the information became as worthy as the substances and the energy, and the quantity of information decides the success and failure in the society. However, in the 21st century, *the post-IT society* is approaching. In other words, ubiquitous society for every body, is going to be realized, where everybody can access to huge amount of information anywhere and anytime. If such post-IT society actually realized, the value of information itself will be depreciated, because huge amount of information can be shared by everybody. The success and failure in the post-IT society depends on whether one can extract

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 30–39, 2008.

essential or useful information or knowledge from massive data. Therefore, in the post-IT society, the development of the methods and technologies for knowledge discovery and knowledge creation are very important.

Informationization also strongly influenced society. According to P. E. Drucker (1993), the capitalism has been moved to the post-capitalist society shortly after the World War II, due to the productivity revolution. This is because, the knowledge became the real, controlling resource and the absolutely decisive factor of production. According to him, *the means of production is no longer capital, nor land, nor labor. It is and will be knowledge.*

1.2 Expansion of Research Object and Change in Scientific Methodology Due to Informationization

The scientific research until the 19th century has developed basically under Newton-Descartes paradigm based on a mechanic view of the world. In the deductive approach, or in theoretical sciences, mathematics played an important role as the language of the science. However, the theory of evolution advocated by C. Darwin in mid 19th century means that every creature in real world evolves and changes with time.

Motivated by such changes of view of real world, in 1891, K. Pearson declared that everything in the real world can be an object of scientific research, and advocated *the grammar of science* (Tsubaki (2002)). It can be considered that the descriptive statistics and subsequent inferential statistics have developed as methodologies of achieving the grammar of science. By the establishment of the method of experimental sciences, not only biology but also many stochastic phenomena in real world such as economy and psychology, became the objects of scientific research.

In the latter half of the 20th century, the computation ability has increased rapidly by the development of the computers. As a result, numerical computation and Monte Carlo computation are applied to the nonlinear dynamics, complex systems, and intertwined high degree of freedom systems that have been difficult to handle by conventional analytic approach based on theoretical science, and the computational science has developed very rapidly.

However, development of IT became a trigger of another development of utilizing the information in rapidly exploding cyber-world. The development of the information technology resulted in accumulation of large-scale massive data in various fields of scientific researches and society, and a huge cyber-world is being created. It is not possible to talk about future development of the science and the technology without establishing methods of effective use of large-scale data. In this article, the scientific methodology supported by the technology of utilizing large-scale data set will be called the "fourth science" (Figure 1).

Needless to say, the first and the second methodologies are the theoretical sciences and the experimental sciences. These sciences are called the deductive method (or principle driven approach) and the inductive method (or data driven approach), and became mainsprings that promoted the scientific researches in the 20th century. However, in the latter half of the 20th century, the computing

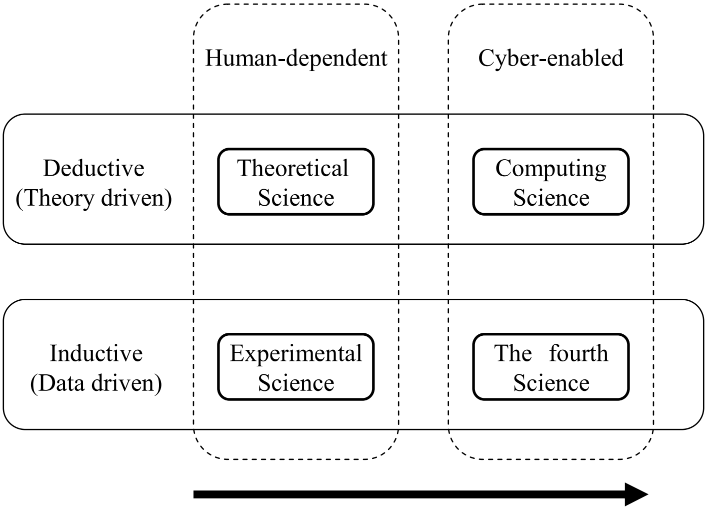

Fig. 1. Four methodologies that drive scientific researches

science was established as a method of alleviating the limit of the theoretical science based on analytic method, and succeeded in the prediction and simulation of nonlinear dynamics, complex systems or intertwined systems.

The computing science and *the fourth science* are newly establishing cyber-enabled deductive and inductive methods while the conventional methodologies, theoretical science and experimental science, relies on the researcher's knowledge and experiences. Now having been developed the computing science, it is indispensable to promote this fourth science strategically to realize well-balanced scientific researches in the information era. It is notable that the U.S. National Science Foundation set "Cyber-enabled discovery and innovation" as a new priority area in the fiscal year 2008 (NSF(2008)).

In the field of global simulation etc., the data assimilation that integrates information obtained from the theoretical model and observations from satellite are becoming popular. In general, this can be considered as a technology to integrate the principle driven approach and the data driven approach. So far, in some area of scientific researches, the integration of two methodologies has been intentionally avoided. However, it is an important subject for the development of the knowledge society in the future. Actually, it can be considered as the filtering method from the standpoints of statistical science or control engineering. A rather natural way of thinking for researchers in methodologies can become the key technology for the science and technology in the future.

The statistics before the inferential statistics, such as the descriptive statistics was based on the observations of the object. On the other hand, the inferential statistics aims at performing scientific reasoning based on carefully designed rather small number of experimental data. However, due to the informationization in recent years, huge amount of heterogeneous data are accumulating, and knowledge discovery from massive large-scale data that are not necessarily designed strictly, became important again. In spite of significant difference of amount of data, it may be said that it is a kind of atavism to descriptive statistics.

In relation to this, Dr. Hotta, President of the Research Organization of Information and Systems, stated an interesting thing about the transition of biology. According to him; "biology that used to be a kind of natural history, became an area of experimental sciences by adopting the scientific methodology in the 20th century. However, now it is becoming possible to decode entire genome of living bodies. In a sense, biology is returning to a kind of natural history in the modern age."

1.3 Active Modeling

In parallel to the changes of the society and expansion of object of the scientific researches, our images of "knowledge" is also radically changing. In the past, a typical definition of the knowledge is "justified true belief," that used to be applied to *being*. However, with the progress of modern age, the knowledge is becoming to applied to *doing* and brought productivity reevaluation and management revolution (Drucker (1993)). Now, approaching to the knowledge society, the knowledge discovery and the knowledge creation are becoming important. Corresponding to these changes, the definition of the knowledge is also changing to "information effective in action and information focused on results."

In the area of statistical science, the role of the model is changing along with the change of the scientific methodology and the image of the knowledge. In the conventional setting of the mathematical statistics, by assuming that the data is obtained from the true distribution, we aimed at performing an objective inference concerning the true structure. However, in the statistical modeling for information processing or for information extraction, it is rather natural to consider that the model is not true or close replica of the truth but is an useful "tool" for information extraction.

Once the statistical model is considered like this, a flexible standpoint of model construction is obtained, namely, in statistical modeling we should use not only the present data but also the theory on that subject, empirical knowledge, and any other data that have been obtained so far, and even the objective of the modeling. Once the model is obtained, the information extraction, knowledge discovery, prediction, simulation, control, and management, etc. can be achieved straightforwardly or deductively (Figure 2). Needless to say, the result of knowledge acquisition using the model leads to refined modeling. In this article, such a process will be called active modeling. Therefore, active modeling forms a spiral of the knowledge development.

To establish the "fourth science" for large-scale massive data, we have the following grand challenges:

1. Prediction and knowledge discovery based on large-scale data,
2. Quantitative risk science, i.e., modeling uncertainty and managing risks,
3. Real world simulation,
4. Service science, i.e., innovations in medical care, pharmacology, marketing, education, etc.

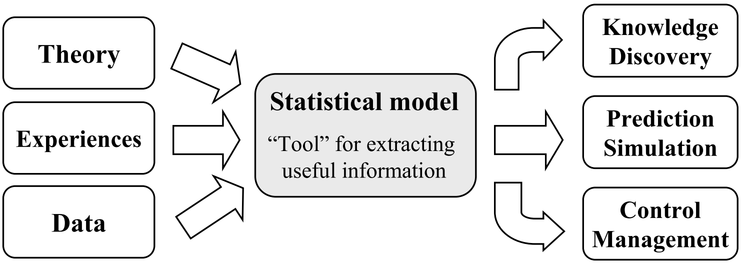

Fig. 2. Active modeling and the use of identified model

In addition, as element technologies for these problems solving, the technologies for the knowledge integration and for the personalization are needed. For the personalization, it is necessary to convert from the formulation of the past statistical inference to an inference about individual object. Of course, this does not mean to abandon the main feature of the statistical inference, namely, the standpoint of capturing stochastic phenomena based on distribution, and can be realized by an appropriate conditioning on the distribution. However, in the modeling for personalization, ultimate conditioning is required, and the difficult problem called "new NP problem" arise, in which the number of variables is much more than the number of observations.

Anyways, the technology that becomes a key to achieve information integration and an ultimate conditioning is the Bayes modeling. It is because various prior information and information from data can be integrated by the use of the Bayes model. Although the Bayes' theorem was discovered middle in the 18th century, and the superiority of the inference based on the Bayes' theorem was well-known, application to real problems was rather rare, due to philosophical controversy, difficulty in determining the prior distributions, and the difficulty in computing the posterior distribution, etc. However, owing to the development of statistical science such as the change in the viewpoint of modeling, the model evaluation criterion that objectively evaluates models that are introduced subjectively, and the development of statistical computing methods such as MCMC and sequential Monte Carlo methods (Kitagawa and Gersch (1996), Doucet et al. (2001)), now the Bayes method becomes the main tool in information extraction, information integration, and information retrieval, etc. (Higuchi et al. (2007)).

Although the Bayes modeling is becoming of practical use, there is one difficulty in the achievement of modeling. Namely, there is no established methodology to derive appropriate class of models for particular problem. Therefore, the researcher's art is still demanded in the most important part of statistical modeling, i.e., the presentation of the model family. The raison d'etre of the researchers, in particular of the statisticians in a cyber world can be found here.

2 Active Modeling of Time Series: Some Examples

So far, we have discussed importance of integrating various kind of information considering the characteristics of the object and objective of the modeling. In this section, we shall show several examples of time series modeling. In time series analysis, the nonlinear non-Gaussian state-space model

$$x_n = f(x_{n-1}, v_n), \quad y_n = h(x_n, w_n) \tag{1}$$

is a useful tool for information extraction and information integration (Kitagawa and Gersch (1996)). Here, y_n, x_n, v_n and w_n are time series, unknown state vector, system noise and observation noise, respectively. The functions $f(x,v)$ and $h(x,w)$ are, in general, nonlinear functions and the distributions of v_n and w_n are not necessarily Gaussian. This general state-space model is a powerful platform for integrating various types of information in time series analysis.

The ordinary state-space model used to be popular in time series modeling because of the presence of the computationally efficient Kalman filter. However, development of sequential Monte Carlo methods for filtering and smoothing with general state-space model opened the door to flexible nonlinear non-Gaussian modeling of time series (Kitagawa and Gersch (1996), Doucet et al. (2001)).

2.1 Prediction and Interpolation by Time Series Modeling

Figure 3 shows the results of increasing horizon prediction of BLSALLFOOD data, the number of food industry workers in US (Kitagawa and Gersch (1996)), based on autoregressive (AR) models with various orders. This time series has apparent seasonal component. In the following prediction, the AR models are estimated based on the first 110 observations and predict the succeeding 46 observations, $y_n, n = 111, \ldots, 156$.

The upper left plot shows the case when AR model with order 1, hereafter denoted as AR(1), was fitted by the Yule-Walker method and obtain the increasing horizon predictive distributions by the Kalman filter. The smooth curve shows the predicted values, i.e. the means of the increasing horizon predictive distributions, and two dotted curves above and below this mean function are the ±1 standard error interval. Almost all actual observations are in these bounds that suggests slight over estimation of the prediction variance. Except for this problem, the prediction looks reasonable. However, the seasonal pattern is not considered at all and this cannot be a good prediction.

The upper right plot shows the results by the AR(3). In this case, the first one cycle was reasonably predicted. But the prediction over one year period is too smooth and is almost the same as the AR(1). The lower left plot shows the results by AR(11). In this case, cyclic behavior was predicted in entire period of four years and the prediction errors are significantly reduced. On the other hand, the lower right plot shows the results by AR(15), the minimum AIC model. Comparing with the prediction by AR(11), it can be seen that much more precise prediction was attained by this model. Actually, it is remarkable that details of the seasonal pattern were successfully predicted by this model.

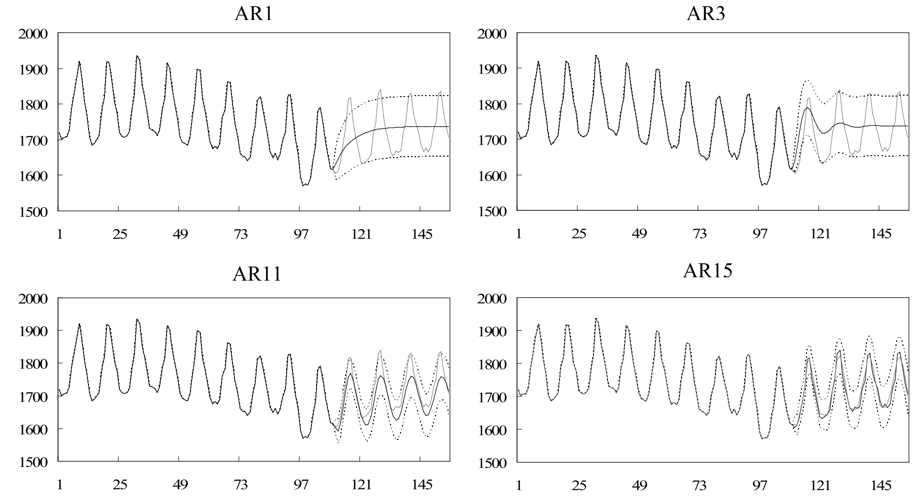

Fig. 3. Increasing horizon prediction of BLSALLFOOD data by AR models with various orders (Kitagawa (2005))

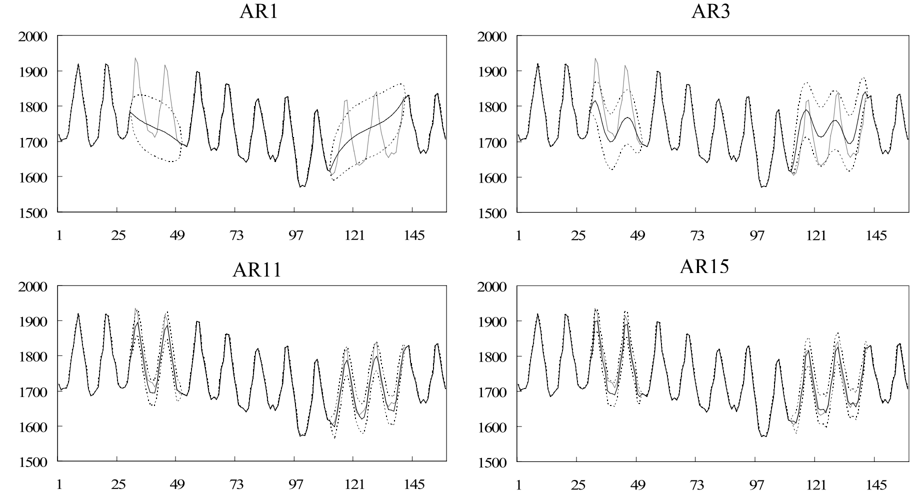

Fig. 4. Interpolation of missing observations by AR models with various orders (Kitagawa and Gersch (1996))

From these results, it can be concluded that even though we obtain the best predictors by the Kalman filter, if the model order is inappropriate, we cannot get good predictive distribution.

Figure 4 show the results of interpolating missing observations by AR models. In this example, 50 observations, $y_{41}, \ldots, y_{60}$ and $y_{111}, \ldots, y_{140}$, are assumed to be missing and are estimated by the fixed interval smoothing algorithm

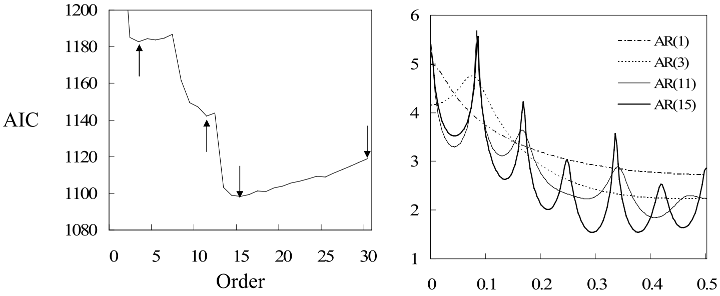

Fig. 5. AIC's and power spectra of the fitted AR models

(Kitagawa and Gersch (1996)). The upper left plot shows the result by AR(1). Approximately 75% of missing observations are included in the ±1 confidence interval. However, similar to the case of increasing horizon prediction, the seasonal pattern of the time series is not used at all for prediction.

The upper right plot shows the result by AR(3). In this case, moderate cyclic pattern was obtained. But the width of the confidence interval is not reduced significantly. In the case of AR(11) shown in bottom left plot, reasonable estimates are obtained by incorporating the annual cyclic pattern. The AR(15) also yields the similar results.

Here, we shall consider from the point of view of the power spectra. The right plot of Figure 5 shows the power spectra obtained by four AR models use for increasing horizon prediction and interpolation. The spectrum by AR(1) is very smooth curve that falls in the right. It can capture the characteristic that the time series has stronger long period components but cannot capture any cyclic behavior. The spectrum by AR(3) has a peak around at $f = 0.08$, corresponding to one year cycle. However, its peak is very broad indicating that it does not capture very definite cyclic pattern. In the case of AR(11), sharp peaks with one year, 6 months and 3 months cycles appeared, but the ones with 4 months and 2.4 months period did not. This means that by AR(11), it is possible to express one year cycle but cannot reproduce every details within the one cycle. On the other hand, in the case of AR(15), any cyclic pattern with one year cycle can be expressed since the spectrum by AR(15) can express 6 periodic components and one direct current component, i.e., $f = 0$.

Incidentally, if we use too higher order models, the spectrum may have more than 6 peaks and it may deteriorate the accuracy of increasing horizon prediction and interpolation. The left plot of Figure 5 shows the values of AIC for various orders of AR models (Konishi and Kitagawa (2007)). The minimum of the AIC was attained at order 15. AR(3) and AR(11) are local minima of AIC but corresponding AIC are significantly larger than that of AR(15). AICfs of the models with order higher than 15 are larger than that of AR(15), suggesting poor prediction abilities than the AR(15).

These results suggest an obvious thing that given a specific model, we cannot breakthrough the limitation of that model. In other words, even if we use the

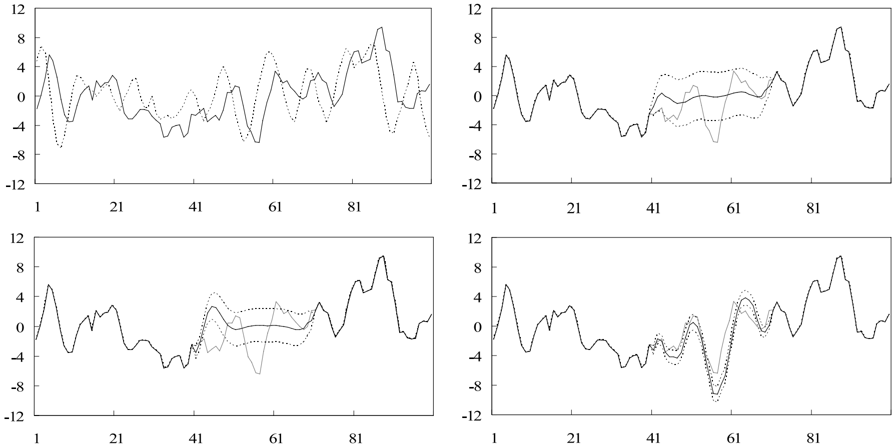

Fig. 6. Interpolation by multivariate AR models

best prediction or best interpolation, it does not guarantee the optimality of the estimation.

2.2 Use of Multivariate Structure

In this subsection, we shall consider interpolation of 2-variate time series (y_n, x_n) and exemplify that the interpolation may significantly improved by incorporating information from other time series. Figure 6 shows an artificially generate 2-variate time series. The problem here is to estimate the data $y_{41}, \ldots, y_{70}$ by assuming that they are missing.

Figure 6 show the result by obtained by using univariate AR model. The best model selected by AIC was AR(5). Since the periodicity is not so strong as the time series of considered in the previous subsection, the interpolated values are very smooth and good reproduction of the missing observations are not achieved even with the AIC best model. This result shows a possible limitation of the univariate time series model for recovering missing data.

To mitigate this limitation, we shall consider the use of information from other time series. Two plots in Figure 7 show the scatter plots of two time series. The left plot show the scatter plot between y_n and x_n, and the right one between y_n and x_{n-2}. Almost no correlation between two time series is seen between y_n and x_n. On the other hand, in the right plot, significant correlation exists between y_n and previous values of other time series, x_{n-2} is seen. These suggest the possibility of improving the prediction or interpolation by using the information about the past values of x_n.

The bottom left plot of Figure 6 show the result obtained by interpolating the missing observations by using the bivariate AR model. It is assumed that on the missing interval both of y_n and x_n are not observed. Although the confidence interval is slightly reduced, the estimated values are similar to those of univariate AR model. On the other hand, the bottom right plot shows the case when we

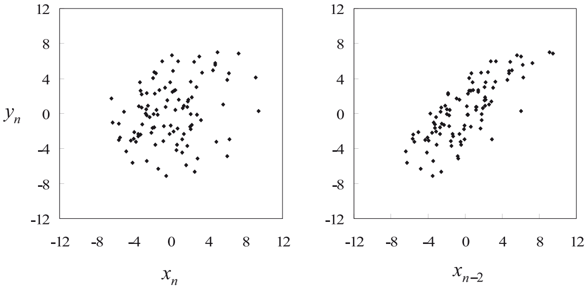

Fig. 7. Scatter plots of two time series

can use the observations of the time series x_n on this interval. Even though we used the same bivariate AR model, very good reproduction of the missing observations of y_n is achieved by using the information of x_n. Although, it is rather obvious, this example clearly shows that we should utilize the all available information appropriate for the current purpose.

References

1. Doucet, A., Freitas, F., Gordon, N.: Sequential Monte Carlo Methods in Practice. Springer, New York (2001)
2. Drucker, P.E.: Post-Capitalist Society. Harper Business, New York (1993)
3. Kitagawa, G., Gersch, W.: Smoothness Priors Analysis of Time Series. Springer, New York (1996)
4. Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer, New York (2007)
5. Matsumoto, N., Kitagawa, G., Roeloffs, E.A.: Hydrological response to earthquake in Haibara well, central Japan–1. Geophysical Journal International 155(3), 885–898 (2003)
6. NSF, NSF wide investment (2008), `http://www.nsf.gov/news/priority_areas/index.jsp`
7. Tsubaki, H.: Statistical science aspects of business (in Japanese). In: Proceeding of the Japan Society of Applied Science, vol. 16, pp. 26–30 (2002)