Automatic Transaction of Signal via Statistical Modeling

Genshiro KITAGAWA and Tomoyuki HIGUCHI

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569 JAPAN

{kitagawa, higuchi}@ism.ac.jp

Received 30 August 1999

Abstract The statistical information processing can be characterized by the likelihood function defined by giving an explicit form for an approximation to the true distribution. This mathematical representation, which is usually called a model, is built based on not only the current data but also prior knowledge on the object and the objective of the analysis. Akaike[2, 3] showed that the log-likelihood can be considered as an estimate of the Kullback-Leibler (K-L) information which measures the similarity between the predictive distribution of the model and the true distribution. Akaike information criterion (AIC) is an estimate of the K-L information and makes it possible to evaluate and compare the goodness of many models objectively. In consequence, the minimum AIC procedure allows us to develop automatic modeling and signal extraction procedures. In this article, we give a simple explanation of statistical modeling based on the AIC and demonstrate four examples of applying the minimum AIC procedure to an automatic transaction of signals observed in the earth sciences.

§1 Introduction: Statistical Modeling

In statistical information processing, a model is built based on not only the current data but also prior knowledge on the object and the objective of the analysis, whereas the conventional data analysis techniques rely on simple manipulation of the current data. To use a proper model for describing the data

makes it possible to combine various knowledge on the object or the information from other data sets, and can enhance a scientific return from the given data sets. Namely, necessary information is extracted based on the model. This is the main feature of statistical information processing.

On the other hand, there is a danger of extracting biased result if an analysis is made by using improper models. Therefore, in information processing based on a model, use of proper model is crucial. Further for an automatic statistical information processing procedures, the development of an automatic statistical modeling procedure is necessary. Akaike information criterion AIC[2] is an objective criterion to evaluate the goodness of fit of statistical model and facilitates the development of automatic statistical information processing procedures.

In this paper, we first briefly review the statistical modeling procedure based on information criterion. Then we shall show examples of developing statistical information processing for knowledge discovery in various fields of earth science.

§2 Information Criterion and Automatic Selection of Models

The phenomena in real world are usually very complicated and information obtained from the real world is in general incomplete and insufficient. The models which we obtained from and used for such incomplete information is inevitably an approximation to the real world. In modeling, it is required to describe the complex real world as precise as possible by simpler model. However, if the objective of the modeling is to obtain precise description of the data, it is not obvious why the model should be simple. A clear answer to this basic question was given by Akaike[2] from a predictive point of view.

In prediction an inference is made on the future data based on the existing data. In statistical prediction, a model is used for prediction and it controls the accuracy of the prediction. If, in the modeling, we adhere to mimic the current data or the phenomenon, then the model will become increasingly more complicated to reproduce the details of the current data. However, by aiming at the improvement of predictive ability, it becomes possible to extract essential information from or knowledge about the object, properly excluding random effects.

Akaike[2, 3] proposed to evaluate the goodness of statistical models by the

goodness of the corresponding predictive distributions. Namely, he proposed to evaluate the goodness of the statistical models by the similarity between the predictive distribution of the model and the true distribution that generates the data $Y_N = [y_1, \ldots, y_N]$, and to evaluate its similarity by the Kullback-Leibler information quantity. Here N is the number of data. Under the situation that the true distribution is unknown, it is not possible to compute the Kullback-Leibler information. However, Akaike showed that the log-likelihood

$$\ell(\theta_m) = \log f_m(Y_N|\theta_m) \tag{1}$$

$$= \sum_{n=1}^{N} \log f_m(y_n|Y_{n-1}, \theta_m), \tag{2}$$

that has been used for many years as general criterion for the estimation of parametric models, can be considered as an estimate of the K-L information (precisely, the expected log-likelihood). Here f_m is one of a set of candidate models for a probability density function of the observation, $\{f_m; m = 1, \ldots, M\}$, which is an approximation to the true distribution, and θ_m is the parameter vector of the density f_m. In particular case where y_n is independently and identically distributed, (2) can be given as a very simple form

$$\ell(\theta_m) = \sum_{n=1}^{N} \log f_m(y_n|\theta_m). \tag{3}$$

An optimal parameter estimate, $\widehat{\theta}_m$, is defined by maximizing the log-likelihood function with respect to θ_m.

According to this idea, the maximum likelihood method can be interpreted as an estimation method that aims at minimizing the K-L information. A difficulty arose in the development of automatic modeling procedures, where the log-likelihoods of the models with parameters estimated from data have biases as estimators of the K-L information, and thus the goodness of the models with estimated parameters cannot be compared with this criterion. This bias occurred because the same data set was used twice for the estimation of parameters and for the estimation of the K-L information. Akaike evaluated the bias of the log-likelihood, and defined the information criterion

$$\begin{aligned} \mathrm{AIC}_m &= -2(\text{log likelihood}) + 2(\text{number of parameters}) \\ &= -2 \log f_m(Y_N|\widehat{\theta}_m) + 2 \|\theta_m\| \end{aligned} \tag{4}$$

by compensating this bias. Here $\|\theta_m\|$ denotes the dimension of the parameter vector. By the use of this AIC, it becomes possible to evaluate and compare the goodness of many models objectively and it enables us to select the best model among many competing candidates $f_m(\cdot|\theta_m)$; $m = 1, \ldots, M$. As a result, the minimum AIC procedure allows us to develop automatic modeling and signal extraction procedures. This is a breakthrough in statistics and helped the change of statistical paradigm from the estimation within the given stochastic structure to modeling with unknown structure. Using AIC, various data structure search procedures and data screening procedures were developed (see e.g., [1, 4, 13]).

As mentioned above, a statistical approach to automatic transaction of data relies on using the minimum AIC procedure which is based on the maximum log-likelihood principle. Then the statistical information processing can be characterized by using the likelihood function defined by giving an explicit form for an approximation to the true distribution from which the data are generated. Although the direct application of the AIC is limited to the model with parameter estimated by the maximum likelihood method, its idea can be applied to much wider class of models and estimation procedures and various types of information criteria are developed recently (e.g.,[7, 11]).

§3 Least Squares Fit of Regression Models

One of the easiest way to represent random effects in the observation y_n is to adopt an observation model in which an observation error (noise) is assumed to be added to a signal $g(n|\theta)$

$$y_n = g(n|\theta) + \varepsilon_n, \quad \varepsilon_n \sim N(0, \sigma^2), \tag{5}$$

where ε_n is an independently and identically distributed (i.i.d.) Gaussian white noise sequences with mean 0 and unknown variance σ^2. In this case, the log-likelihood can be given by (3) and the maximum likelihood estimates is obtained by minimizing $\sum_{i=1}^{N} [y_n - g(n|\theta)]^2$. Within this framework, the major efforts in developing an automatic procedure are made on preparing a wide variety of candidates for $g(\cdot|\theta)$. We shall show two examples. In each case the determination of the best $g(\cdot|\theta)$ through the AIC plays an important role in making the procedure automatic and objective.

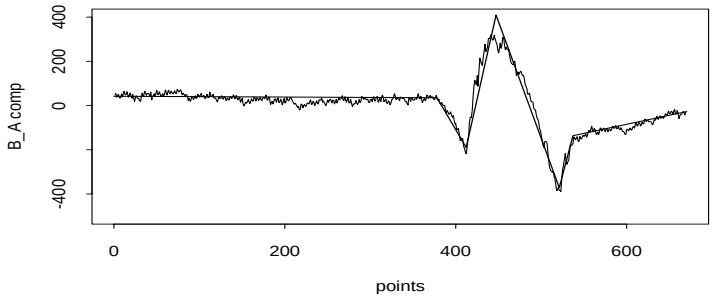

Fig. 1 Magnetic field perturbation associated with the LSFAC and fitted polyline.

3.1 Automatic Identification of Large-Scale Filed-Aligned Current Structure

The plasma stream from the Sun, solar wind, interacts the Earth's magnetic field and generates three-dimensional current system above the ionosphere. Because conductance along the magnetic field is much higher than that across the magnetic field, the currents flow along magnetic field lines. Such current is called the large-scale field-aligned currents (LSFAC). LSFACs are also related to the dynamics of aurorae. Depending on the number of LSFAC sheets crossed by a satellite and also on the intensity and flow direction (upward/downward) of each LSFAC, a plot of the magnetic fluctuations associated with the LSFAC (as shown in Fig. 1), mainly in the east-west (E-W) magnetic component, can have any shape, and we have been depending on visual examination to identify LSFAC systems. We developed a procedure to automatically identify the spatial structure of LSFAC from satellite magnetic field measurements.[5)]

The required task is to automatically fit the first-order B-spline function with variable node positions, which is sometimes called a polyline or linear spline.[8)] Namely, we adopt a polyline as $g(n|\theta)$ mentioned above. Although node points are fixed in usual spline applications, the benefit of the spline function can be maximized when node points are allowed to move.[6)] We therefore treat a set of node positions and node values as parameters to be estimated. In addition, the number of node points, which determines the number of LSFAC sheets, is one of the fitting parameters. For this modeling, the AIC with J node points is defined by

$$\mathrm{AIC}_J = N \log \widehat{\sigma}_J^2 + 2(2J+1) + \text{constant}, \tag{6}$$

where $\widehat{\sigma}_J^2$ is the maximum likelihood estimates of the variance of the observation error in (5).

We applied the developed procedure to the whole data set of magnetic field measurements made by the Defense Meteorological Satellite Program–F7 (DMSP–F7) satellite during the entire interval of its mission from December 1983 to January 1988. DMSP is a Sun-synchronous satellite with a nearly circular polar orbit at about 835 km in altitude, and thus the orbital period is about 101 minutes. We divide a data file of each polar pass into two parts, dayside and nightside files, by the data point of the highest-latitude satellite position. We have a total of 71,594 data files. Each data file usually contains from 600 to 800 magnetic field vector measurements as well as various geographic and geomagnetic parameters necessary for describing a satellite position at observation time. The sampling interval is 1 second.

The first subject made possible by the developed procedure is to find a four-FAC-sheet structure along dayside passes. Ohtani *et al.* [12] reported only four events observed by the DMSP–F7. This four-FAC-sheet structure was unexpected phenomena from a viewpoint of the conventional interpretation of the LSFAC, and happened to be discovered. The developed procedure found 517 northern and 436 southern passes along which the DMSP–F7 observed four LSFACs. This discovery allowed us for the first time ever to conduct a statistical study on what solar wind conditions bring about this peculiar LSFAC. In addition, the developed automatic procedure to identify the structure of LSFAC systems can be used to conduct space weather forecasting that is becoming an important subject in space science, as space environment, because it is influential to the operation of satellites, and more relevant to human activities.

3.2 Automatic Determination of Arrival Time of Seismic Signal

When an earthquake occurs, its location is estimated from arrival times of the seismic waves at several different observatories. In Japan, it is necessary to determine it very quickly to evaluate the possibility of causing Tsunami. Therefore, the development of computationally efficient on-line method for automatic estimation of the arrival time of the seismic wave is a very important problem. At each observatory, three-component seismogram is observed at a sampling interval of about 0.01 second.

When seismic wave arrives, the characteristics of the record of seismograms, such as the variances and the spectrum, change significantly. For estimation of the arrival time of the seismic signal, it is assumed that each of the seismogram before and after the arrival of the seismic wave is stationary and can be expressed by an autoregressive model as follows [14]:

Background model: $$y_n = \sum_{i=1}^{m} a_i y_{n-i} + v_n, \quad v_n \sim N(0, \tau_m^2)$$

Seismic signal model: $$y_n = \sum_{i=1}^{\ell} b_i y_{n-i} + w_n, \quad w_n \sim N(0, \sigma_\ell^2).$$

Although the likelihood function of y_n for the AR modeling depends on Y_{n-1}, we can obtain its analytic form as a function of the AR coefficients. Then, given the observations, AIC of the locally stationary AR model is obtained by

$$\mathrm{AIC}_k = k \log \widehat{\tau}_m^2 + (N-k) \log \widehat{\sigma}_\ell^2 + 2(m+\ell+2), \tag{7}$$

where N and k are the number of data and the assumed arrival time point, and $\widehat{\tau}_m^2$ and $\widehat{\sigma}_\ell^2$ are the maximum likelihood estimates of the innovation variances of the background noise model and the seismic signal model, respectively. In this locally stationary AR modeling, the arrival time of the seismic wave corresponds to the change point of the autoregressive model. The arrival time of the seismic signal can be determined automatically by finding the minimum of the AIC_k on a specified interval.

However, for automatic determination of the change point by the minimum AIC procedure, we have to fit and compare $K \times (M+1)^2$ models. Here K is the number of possible change points and M is the possible maximum AR order of background and signal models. The fitting and finding the minimum AIC model can be realized computationally efficiently by using the least squares method based on the Householder transformation. By this method, the necessary amount of computation is only twice as much as that for the fitting of single AR model with order M.

Fig. 2A shows a portion of a seismogram of a foreshock of Urakawa-Oki Earthquake observed at Moyori, Hokkaido, Japan. Fig. 2B shows AIC_k for $k = 850, \ldots, 1150$. From this figure, it can be seen that the AIC becomes the minimum at $k = 1026$. Using the estimated arrival times of seismic signal at several observatories, it is possible to estimate the epicenter of the earthquake automatically. Also, the AIC values shown in Fig. 2B can define the likelihood function for the arrival time. Based on that function, it is expected to be able

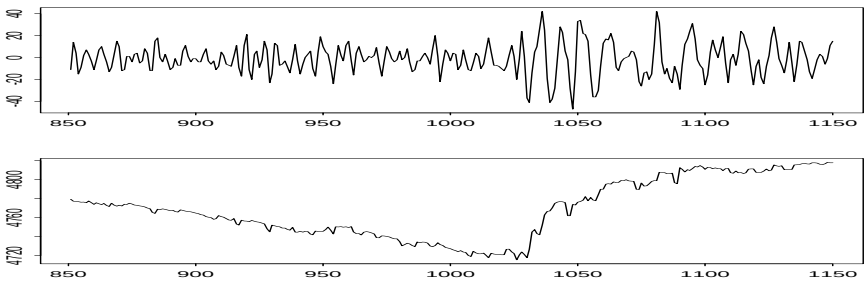

Fig. 2 A: Seismic signal. B: AIC value.

to develop a new maximum likelihood type estimator for the epicenter of the earthquake.

§4 Generalized State Space Model

The two examples shown above are based on the relatively simpler models which can produce an analytic form of the AIC, because of the simple assumptions such that the observation error is an i.i.d. Gaussian white noise sequence and the regression model is linear. Although the adopted models are suited to each application, general framework for describing the models is available for time series data. This framework is called the generalized state space model which is a generalization of the state space model,[9] and is defined by

$$x_n \sim q(\cdot|x_{n-1},\theta) \qquad \text{[system model]} \tag{8}$$

$$y_n \sim r(\cdot|x_n,\theta) \qquad \text{[observation model]} \tag{9}$$

where x_n is the state vector at time n. q and r are conditional distributions of x_n given x_{n-1} and of y_n given x_n, respectively. This generalized state space model can treat the non-Gaussian and non-linear time series model, in contrast to the ordinary state space model.

For state estimation in the generalized state space model, one step ahead predictor and filter can be obtained by the following recursive formulas called the non-Gaussian filter:

[prediction]

$$p(x_n|Y_{n-1}) = \int_{-\infty}^{\infty} p(x_n|x_{n-1})p(x_{n-1}|Y_{n-1})dx_{n-1} \tag{10}$$

[filtering]

$$p(x_n|Y_n) = \frac{p(y_n|x_n)p(x_n|Y_{n-1})}{p(y_n|Y_{n-1})}, \tag{11}$$

where we omit a dependency on θ for simple notation. $p(y_n|Y_{n-1})$ in the filtering is obtained by $\int p(y_n|x_n)p(x_n|Y_{n-1})dx_n$ and then the log-likelihood in the generalized state space model can be defined by (2).

4.1 Automatic Data Cleaning

In an attempt to predict big earthquake anticipated in Tokai area, Japan, various types of measurement devices have been set since 1979. The underground water level is observed in many observation wells at a sampling interval of 2 minutes for 20 years. However, the actual underground water level data contains huge amount of (1 % to over 10 %, depending on the year) missing and outlying observations. Therefore, without proper cleaning procedure, it is difficult to fully utilize the information contained in the huge amount of data. We interpolated the missing observations and corrected the outliers by using a non-Gaussian state space model:[10)]

$$t_n = t_{n-1} + w_n, \qquad y_n = t_n + \varepsilon_n, \tag{12}$$

where $w_n \sim N(0, \tau^2)$. For the observation noise ε_n, we considered Gaussian mixture distribution

$$\varepsilon_n \sim (1-\alpha)N(0, \sigma_0^2) + \alpha N(\mu, \sigma_1^2), \tag{13}$$

where α is the rate of contaminated observations, σ_0^2 the variance of ordinary observations and μ and σ_1^2 are the mean and variance of the outliers. Such a density allows the occurrence of large deviations with a low probability. In this model, a set of $[\tau^2, \alpha, \sigma_0^2, \mu, \sigma_1^2]$ is the parameter vector θ to be optimized. For the filtering and smoothing of the non-Gaussian state space model, we applied the non-Gaussian filter and smoother [9)]. By this Gaussian-mixture modeling of the observation noise, the essential signal t_n is extracted automatically taking account of the effect of the outliers and filling in the missing observations.

4.2 Finding out the Effect of Earthquake in Underground Water Level Data

Even after filling in the missing observations and correcting the outliers, the underground water level is very variable. Further, because the data is affected by many other covariates such as barometric air pressure, earth tide and precipitation, it is almost impossible to extract the effect of earthquake by simple manipulation of the data. In an attempt to account for the effect of the covariates on the underground water level, we considered the following model,

$$y_n = t_n + P_n + E_n + R_n + \varepsilon_n, \tag{14}$$

where t_n, P_n, E_n, R_n and ε_n are the trend, the barometric pressure effect, the earth tide effect, the rainfall effect and the observation noise components, respectively.[10] We assumed that those components follow the models

$$\nabla^k t_n = w_n, \qquad P_n = \sum_{i=0}^{m} a_i p_{n-i}, \tag{15}$$

$$E_n = \sum_{i=0}^{\ell} b_i et_{n-i}, \quad R_n = \sum_{i=1}^{k} c_i R_{n-i} + \sum_{i=1}^{k} d_i r_{n-i} + v_n.$$

Here p_n, et_n and r_n are the observed barometric pressure, the earth tide and the observed precipitation at time n, respectively. These components can be estimated by the state space representation of (8) and (9) and by the use of Kalman filter and the fixed interval smoother.

Fig. 3B-E show the extracted coseismic effect, air pressure effect, earth tide effect and precipitation effect obtained by the Kalman smoother. The annual variation of the trend is only about 6cm and the effect of the earthquake with magnitude M=4.8, at a distance D=42 km, is clearly detected. Most of the range of about 45 cm trend variations in Fig. 3A can be considered as the effect of barometric pressure, etc.

Fig. 4 shows the scatter plot of the earthquakes with log-distance as the horizontal axis and the magnitude as the vertical axis. The earthquakes with ◯ -label have detected coseismic effects over 8 cm. Whereas the box and △-labeled events indicate the earthquakes with coseismic effects over 4 and 1 cm. The + labeled events indicate earthquakes without coseismic effects over 1 cm. Two lines in the figure are defined by

$$\bar{M} = M - 2.62 \log_{10} D = C$$

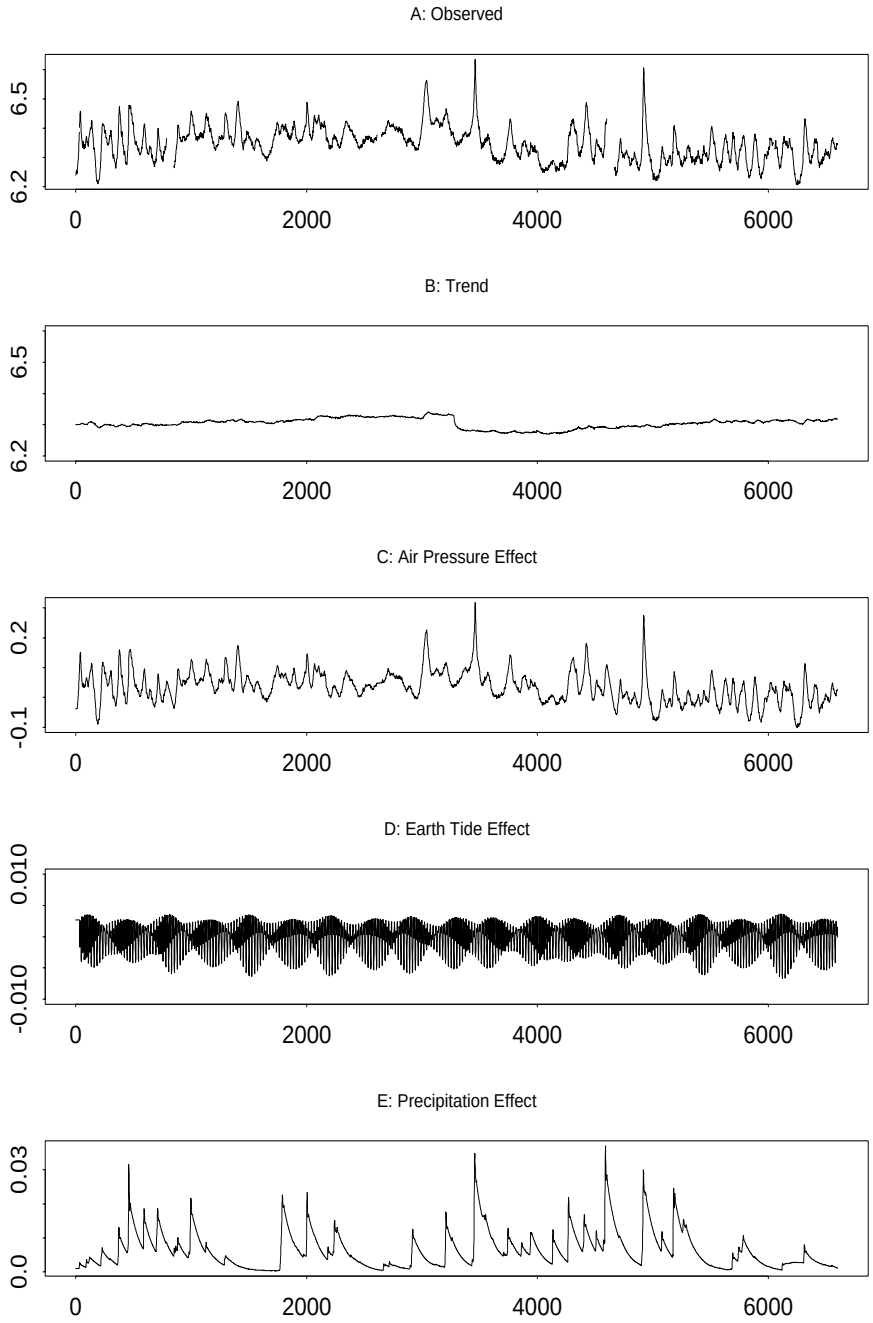

Fig. 3 A: A segment of the water level data. B: The extracted seismic effect.

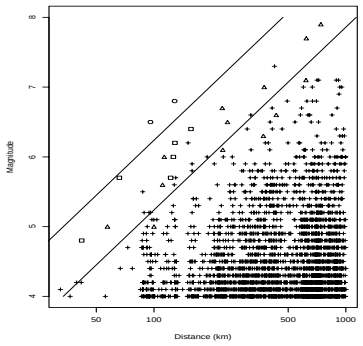

Fig. 4 Scatter plot of the earthquakes.[10)]

for $C = 0$ and 1.

From the analysis of over 10 years data, we obtained the following important findings: (1) The drop of water level can be seen for most of the earthquakes with magnitude larger than $M > 2.62 \log_{10} D + 0.2$, where D is the hypocentral distance. (2) The amount of the drop can be explained as a function of $M - 2.62 \log_{10} D$. (3) Except for the coseismic effect drop, the trend regularly increases at the rate of about $6cm$ per year.

§5 Summary

In statistical approach to knowledge discovery, a proper modeling of the object is a crucial step. In this paper, we introduced an automatic procedure based on Akaike information criterion, AIC. We demonstrated four applications of the minimum AIC procedure to the actual large data sets observed in the earth science. In each case, an appropriate representation of the signals enables us to give an explicit form of the AIC and results in the realization of an automatic procedure to handle a large amount of data sets.

Acknowledgment One of the author (G.K.) thanks Prof. Tetsuo Takanami of Hokkaido University, Japan and Dr. Norio Matsumoto of the Geological Survey of Japan for the cooperative works related to Section 3.2, 4.1, and 4.2. The other (T.H.) thanks Dr. Shin-ichi Ohtani of The Johns Hopkins University, Applied Physics Laboratory for his useful suggestions in improving the procedure explained in Section 3.1.

References

1) Akaike, H., "Automatic Data Structure Search by the Maximum Likelihood", in *Computers in Biomedicine, A Supplement to the Proceedings of the Fifth Hawaii International Conference on Systems Sciences*, Western Periodicals, California, pp. 99–101, 1972.

2) Akaike, H., "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium in Information Theory* (Petrov, B.N. and Csaki, F. eds.), Akademiai Kiado, Budapest, pp. 267–281, 1973. (Reproduced in *Breakthroughs in Statistics, Vol.I, Foundations and Basic Theory* (S. Kots and N.L. Johnson eds.) Springer-Verlag, New York, 610–624, 1992.)

3) Akaike, H., "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control, AC-19*, 716–723, 1974.

4) Bozdogan, H. ed. *Proceeding of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, Kluwer Academic Publishers, 1994.

5) Higuchi, T. and Ohtani, S., "Automatic Identification of Large-Scale Field-Aligned Current Structure," *Research Memorandom*, No. 668, The Institute of Statistical Mathematics, 1998. `http://www.ism.ac.jp/~ higuchi/AIFACpaper.html`

6) Hiragi, Y., Urakawa, H., and Tanabe, K., "Statistical Procedure for Deconvoluting Experimental Data," *J. Applied Physics, 58, 1*, 5–11, 1985.

7) Ishiguro, M., Sakamoto, Y. and Kitagawa, G., "Bootstrapping log likelihood and EIC, an extension of AIC," *Annals of the Institute of Statistical Mathematics, 49, 3,* 411–434, 1997.

8) Ja-Yong, Koo, "Spline Estimation of Discontinuous Regression Functions," *J. Computational and Graphical Statistics, 6, 3,* 266–284, 1997.

9) Kitagawa, G. and Gersch, W., *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statistics, No. 116, Springer-Verlag, New York, 1996.

10) Kitagawa, G. and Matsumoto, N., "Detection of Coseismic Changes of Underground water Level," *Journal of the American Statistical Association, 91, 434,* 521–528, 1996.

11) Konishi, S. and Kitagawa, G., "Generalized Information Criteria in Model Selection ", *Biometrika, 83, 4,* 875–890, 1996.

12) Ohtani, S., Potemra, T.A., Newell, P.T., Zanetti, L.J., Iijima, T., Watanabe, M., Blomberg, L.G., Elphinstone, R.D., Murphree, J.S., Yamauchi, M., and Woch, J.G., "Four large-scale field-aligned current systems in the dayside high-latitude region," *J. Geophysical Research, 100, A1,* 137–153, 1995.

13) Sakamoto, Y. and Akaike, H., "Analysis of cross-classified data by AIC," *Annals of the Institute of Statistical Mathematics, 30, 1,* 185–197, 1978.

14) Takanami, T. and Kitagawa, G., "Estimation of the Arrival Times of Seismic Waves by Multivariate Time Series Model", *Annals of the Institute of Statistical mathematics, 43, 3,* 407–433, 1991.