Information processing based on hierarchical time series models: Bayesian self-tuning and parallel computing

Tomoyuki Higuchi[1]

[1] Inst. Stat. Math.

http://www.ism.ac.jp/higuchi/

A hierarchical structure of the statistical models involving the parametric, state space, generalized state space, and self-organizing state space models is explained. It is shown that by considering higher level modeling, it is possible to develop models quite freely and then to extract essential information from data which has been difficult to obtain due to the use of restricted models. It is also shown that by rising the level of the model, the model selection procedure which has been realized with human expertise can be performed automatically and thus the automatic processing of huge time series data becomes realistic. In other words, the hierarchical statistical modeling facilitates both automatic processing of massive time series data and a new method for knowledge discovery.

Data always contain some errors and thus only by proper processing of the errors, it becomes possible to separate the essential or universal part from the errors that occurred only for that data. However, in the analysis of complex, massive and/or multivariate phenomena with nonstationarity or nonlinearity, it is almost impossible to express it from simple scientific theory and consider the difference from the actual data as observation noises. In such a case, by reasonably decomposing the data into a part that can be explained from the existing knowledge and other, the possibility of new scientific discovery emerges.

In the frontiers of sciences, unexpected phenomena appear at the very limit of the errors. Therefore, by simply considering the unknown part as the observation errors, it is almost impossible to find out the clue to the discovery. For the discovery in the frontiers of sciences, it is crucially important to express our expectation on the unknown part as a form of model, and perform the extraction of the information very actively.

Needless to say, in such a statistical modeling, use of appropriate model is crucially important. If the data did not contain any errors, or the objective of our analysis was just to describe the phenomena precisely, it is sufficient to use the model with highest ability of description. However, in the actual analysis our objective is often to extract or discover a more universal knowledge. In the statistical science, this essential part is considered from the predictive point of view.

To obtain good models, it is necessary to develop appropriate model class and model evaluation criterion. Further, to make the modeling practical, it is also necessary to develop efficient computational method. In general, flexible models with high ability of description inevitably contain increasingly many unknown parameters and require huge amount of computations for the estimation of them. In this study, we will consider the modeling of the time series which is the most important in the statistical analysis of massive data. For flexible modeling of time series, it is necessary to use a model with the number of parameters proportional to the data length N. For such models, the ordinary computational methods requires the computation with order O(N^3), and are obviously unrealistic to apply. Without skillful computational methods which are based on the essential mathematical structure of the model, it is sometimes impossible to obtain reasonable estimates of the unknowns. Further, to treat massive data based on sophisticated models, it is essential to develop a computationally efficient method.

In the case of time series model, due to a mathematical structure, i.e., Markov property, a large class of models can be expressed by the generalized state space model. This generalized state space model has another significant merit that it automatically realizes the estimation of its unknowns with O(N) computations by the recursive filter and smoothing algorithms. Therefore, the generalized state space model is a useful platform for flexible modeling and at the same time is a base for generating efficient computation.

In this study, it will be shown that by considering higher level modeling, it is possible to develop models quite freely. By a numerical example, it will be shown that, as a result, it becomes possible to extract essential information from data which has been difficult due to the use of restricted models. It was also shown that by rising the level of the model, the model selection procedure which has been realized with human expertise can be performed automatically and thus the automatic processing of huge time series data becomes realistic. In other words, the high level statistical modeling facilitates both automatic processing of massive time series data and a new method for knowledge discovery.