Bulletin
Volume 23 Number 4
March 2001

Editor's Foreword

Nowadays, databases can range in size into the terabytes. Within these masses of data lies hidden pattern of strategic importance. But when there are so much data, how can one draw meaningful conclusions? The newest answer is data mining, which finds the hidden gold in the data using sophisticated techniques. Recently, the City University of Hong Kong, in a partnership with the SAS Institute, has set up Hong Kong's first "Knowledge Discovery Centre" with an emphasis on data-mining. In the interview section of this issue, the centre's director, Professor Yer-Van Hui, talks about the importance of data-mining and its relevance to the statistics community.

This issue also features an article by C.S. Wong and W.K. Li of the University of Hong Kong on an application of the mixture time series models. In the President's forum, Professor Li reported the latest development of the Society.

The last issue of this volume brings the end of my two-year period as the Bulletin's editor and Dr. Ping Shing Chan of the Chinese University of Hong Kong will take over this onerous task in June 2001. In addition to the joy of vacating this position, it means that this is my last Editor's Foreword.

Alan Wan

CONTENTS

(Vol. 23/No. 4, March 2001)

President's Forum

Professor W.K. Li

In this belayed March issue of the Bulletin I would just like to thank you all for the support of the 2000-2001 session.

In last year, the Secondary School Statistical Project Competition has continued to flourish and has attracted a record number of entries. We are particular grateful to the SPC organising committee under Cecilia Chan and colleagues who serve as adjudicators. We are, of course, particularly grateful for Hang Seng Bank for its support throughout the years.

As for local examinations and accreditation the negotiation with RSS is reaching its final stage. Mr Goddard, the Secretary of RSS will visit Hong Kong in early June. He will hold meetings with Mr W H Fung, the project co-ordinator and hopefully all the obstacles could be cleared finally. In anticipation of the final clearance, Mr Fung has already formed liaison sub-groups with Institutions in China, local tertiary institutions, vocational institutions and professional bodies.

I think we all agree that the Bulletin plays an important role in disseminating news about the statistical community in Hong Kong and serves as an important place of dialogue among members. In this connection, I would like to thank Dr Alan Wan for his role as Publication Secretary of the Society for three successive years. It is really a job that requires a lot of input and sacrifice in time and Alan has really done a great job. Now that he has decided to take a break from being involved as a council member and I wish him every success in his academic career.

Let's also congratulate Professor Howell Tong for receiving this year's National Natural Science Award (Class II). Professor Tong is the only mathematician over the entire China and the only scholar from Hong Kong SAR to have such an honour this year. It is also the first time that such an award has been given to a statistician. This is all the more significant as there is no class I award (which is seldom given anyway) and there are only 15 class II awards this year. The class III and IV awards have also been abolished from this year. This is clearly a great honour not just to Professor Tong but also to the entire Hong Kong statistical community.

Finally, I look forward to your continued support for the session 2000-2001.

Discovering knowledge from data
- a conversation with Professor Yer-Van Hui

Alan Wan
The City University of Hong Kong

In recent months, the term "data mining" has been on the tip of everyone's tongue at the City University of Hong Kong following the establishment of a "Knowledge Discovery Centre" in the university's Department of Management Sciences late last year. The Centre, directed by Professor Yer-Van Hui, is a partnership between City U. and the SAS Institute with an aim of familiarizing students with the latest internet concepts, applications and analytical techniques. In addition, businesses in Hong Kong and China will benefit by being able to access the university's training and consulting services.

The following interview was conducted on 2 January 2001, in which Professor Hui shared with us his view on data mining and in particular, the relevance of data mining for the statistics profession.

Y.V.: Professor Yer-Van Hui
Alan: Dr. Alan Wan

Alan: *Thanks for sparing the time for this interview. I know that City University has recently set up Hong Kong's first Knowledge Discovery Centre with the SAS Institute. This is a pretty new thing, not only in Hong Kong, but also in the entire Asia-Pacific region. May I start by asking how this idea was originally conceived and the background of the partnership with SAS?*

Y.V.: Yes, in fact, data mining is closely related to statistics. All along statisticians' emphasis has been on the handling and analysis of small data sets. But the I.T. revolution witnessed in the past decade has brought with it a proliferation of data; nowadays, the data sets are much, much bigger, and data warehouses consisting of billions of items of

information are commonplace. The objective of data mining is to unearth the hidden gold in the data and use it to improve the profitability of an organization. In fact, data mining is at the interface between statistics, computer science and business modeling. In our department, all our colleagues have undergone rigorous training in statistics, and linking up with the SAS Institute gives us the benefit of being able to access the most up-to-date e-intelligence software. Also, as part of a business faculty, our research has always had a strong focus on business applications, and through consulting work our team has built up a wealth of knowledge on business modeling. All these characteristics place us in a privileged position for our recent activities on data mining, and it's only natural that we are taking the lead in this work.

Alan: *So what's the linkage between data mining and data warehousing?*

Y.V.: Well, in an e-intelligent environment, the data must be in place before we can embark on any sensible work on data mining. In fact, "data mining" and "data warehousing" are both integral components of an e-intelligent system. In data mining, it's important that we have a problem, a target or a goal in mind before we can proceed onto the stage where we retrieve the information of relevance to the organization. It could be a tricky thing to extract the information because the data might be hidden in various databases, and often the data need to be tidied up, too. So if an organization has a good data-warehouse, i.e., the data are linked by the system, then it is easier to retrieve the data and that can save enormous time and energy. In this sense, data warehousing and data mining are integrally related.

Alan: *Okay. So how is your department adapting itself to its new emphasis on data mining?*

Y.V.: In one way, we are building up our academic strength in data mining and in the broader area of e-intelligence. Through our teaching programs we are equipping students with knowledge in these areas, so that in future our undergraduate students in Managerial Statistics will have all gone through training in data mining and data warehousing before graduating. On the graduate level, we've also introduced courses on data mining for the MBA, MA Quantitative Analysis for Business and E-Commerce students. Other than that, we've also planned to run courses on customer relationship management with data-mining emphasis for our undergraduate Service Operations Management majors. Also, as you know, recently we've teamed up with SAS, and through public seminars, occasional training courses and consulting work, our expertise is being transferred to the wider business community. In fact, it is a reciprocal knowledge transfer process, as by going through the interaction phase we also learn from the business sector its

valuable practical experience. Such knowledge can be brought back to the classroom for teaching and also serves as a stimulus for our research. Indeed, the mission of our centre is to be a centre of knowledge transfer, and it is hoped that the centre can help enhancing Hong Kong's business competitiveness and contribute to Hong Kong's transformation into a knowledge society.

Alan: *It seems to me that data mining tools have been around for nearly a decade. Do you think that Hong Kong is lagging behind other countries in this and the general area of e-intelligence?*

Y.V.: In fact, data mining first started in the U.S. But in Hong Kong, it is still at its infant stage and there's a definite need to train more people in Hong Kong with knowledge in data mining. I think among the Southeast Asian nations, Singapore has taken the lead in this area and Hong Kong is somewhat lagging behind.

Alan: *Would you consider data mining a part of statistical modeling? Or is there any similarity that you can draw between data mining and statistical modeling?*

Y.V.: I think the approaches of the two are somewhat different. In most cases of statistical modeling, we postulate then estimate our models. Eventually based on the Chi-Square test or other model selection criteria, a preferred model is chosen. On the other hand, in data mining, because enormous data are available, so we can afford to use a trial and error process. Often we start with a model and use a segment of data to "try out" the model. In fact, in the jargon of data-mining this is called "learning". In each round of "learning", we use the results to adjust the model, which is then progressively tuned up with more and more data. So this is different from statistical modeling, in which the entire set of data is usually used from the initial stage. In data mining, the "learning" process changes the model each time until an acceptable model comes up. Having said this, data mining is of little use without statistics, because at the end of the day, it is the statistical techniques that choose the final model. If one has no statistical knowledge it is hard to know how the model comes about. Also, how does one deal with the issues of say, missing data or variable transformation? Ultimately one will have to rely on statistics to solve these issues.

Alan: *Steering the conversation now to your own research agenda. Do you think your research will focus on data mining from now on?*

Y.V.: Well, this is not exactly related to what we've been talking about. Speaking of my own research, basically I am a nosy person and I enjoy learning many different things. In fact, I think research; teaching and consulting can be integrally related but also do not necessarily have to be related. I have, for

example, done work on production management and time series, but I have never taught these subjects at universities. Quite similarly, I've been interested in computers ever since my time at graduate school, and I've worked on statistical computing, so it is only natural that I'm developing an interest in data mining as it is at the interface of statistics, computing and business modeling.

Alan: *So given the recent development in data mining, do you think that the subject is becoming an indispensable part of a statistician's training?*

Y.V.: I totally agree with what you said, as all business transactions are done through I.T. nowadays. It's a simple process, and as soon as the transactions are done, the related information goes into the databases. For a large organization we're talking about thousands of receipts every day with dozens of items on the receipts. This means that statisticians must come to grips with having to analyze huge amounts of data, and getting acquainted with data mining techniques becomes a necessity.

Alan: *Okay, then what sorts of facilities are available in Hong Kong if one wishes to get better acquainted with data mining?*

Y.V.: Once in a while the software houses organize training courses, which are in the form of an introductory seminar on data mining or on the software. But as far as I know, no institution has yet offered any training course with an in-depth discussion on data mining. So one thing our centre is planning to do in the future is to organize data mining courses that last for 2-3 days, in which the students will acquire hands-on experience with the techniques using the SAS software.

Alan: *That sounds great. I think that basically sums up what we intended to discuss today. Thanks again for sparing the time for the interview.*

Y.V.: My pleasure.

A mixture time series model for the Hang Seng Index

C. S. Wong and W. K. Li
The University of Hong Kong

In many real life situations we encounter time series data with a multi-modal marginal or conditional distribution. See e.g. Chan and Tong (1998). Traditionally, a natural way to model multi-modal data is by means of mixtures. Recently, Wong and Li (2000) present a detailed study of mixture autoregressive (MAR) models.

A major feature of this type of model is that under regularity conditions one can mix non-stationary and stationary AR components but still obtains a (second order) stationary time series overall.

Another interesting feature is that the MAR model can possess multi-modal predictive densities. This feature may be of particular relevance with regard to the value-at-risk problem in finance.

The parameters of a mixture AR model can be easily estimated using an expectation maximization (EM) algorithm. Model selection can be done by means of a Bayesian information criterion (BIC). Further extensions to incorporate conditional heteroscedasticity and transfer functions can be found in Wong and Li (2001a, b). Here we illustrate the model by applying it to the daily closing Hang Seng Index.

Denote the time series under consideration by y_t. The K-component mixture autoregressive (MAR) model under consideration is defined by

$$F\left(y_t \mid H_{t-1}\right) = \sum_{k=1}^{K} \alpha_k \Phi\left(\frac{y_t - \phi_{k0} - \phi_{k1} y_{t-1} - \ldots - \phi_{kp_k} y_{t-p_k}}{\sigma_k}\right) \qquad (1)$$

We denote this model by MAR(K; p_1, p_2, …, p_k). Here $F(y_t \mid H_{t-1})$ is the conditional cumulative distribution function of Y_t given the past information, evaluating at y_t; H_t is the information set up to time t; $\Phi(\cdot)$ is the (conditional) cumulative distribution function of the standard Gaussian distribution; $\alpha_1 + \ldots + \alpha_K = 1$, $\alpha_k \geq 0$, $k = 1,\ldots,K$.

We consider the HSI data in three periods: the whole year data for the years 1971, 1987 and the indexes from 2 January to 28 November for the year 1997. There are 243, 246 and 223 observations in the years 1971, 1987 and 1997 respectively. We fitted the MAR models to the return series, or the log-differenced series, that is y_t = log(HSI at time t) – log (HSI at time $t - 1$).

The best model for the year 1971 is the MAR(2;1,1) model with $\phi_{k0} = 0$, k = 1,2, namely,

$$F(y_t \mid H_{t-1}) = .4731\Phi\left(\frac{y_t - 1.0128 y_{t-1}}{.0156}\right) + .5269\Phi\left(\frac{y_t - .3616 y_{t-1}}{.0126}\right)$$

The value of BIC is –1712.48. The standard errors of $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\phi}_{11}, \hat{\phi}_{21})$ are (.0833, .0833, .0013, .0011, .1270, .0703). In this model, the estimate of ϕ_{11} is close to one which indicates part of the series being governed by an integrated series of order 2, i.e. I(2). Hence, the logarithm of the original series is governed by a mixing of a I(1) component and a I(2) component.

After examining all the plots of the predictive distribution, we observe that the predictive distribution will be bimodal if the last period return is more than 2% in magnitude. Similar observations are made in the years 1987 and 1997. Some typical predictive distributions generated from the fitted MAR model are shown in Figure 1.

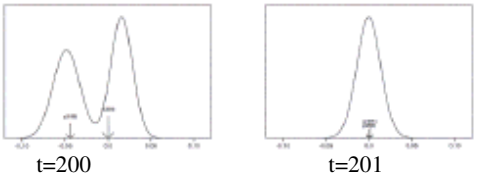

Figure 1: The predictive distribution of the 1971 HSI return series for t = 200 and t = 201. The actual values at t and $t - 1$ are also shown.

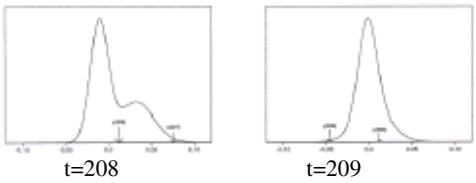

Figure 2: The predictive distribution of the 1987 HSI return series for t = 208 and t = 209. The actual values at t and $t - 1$ are also shown.

For the year 1987, the best model is the MAR(3;1,1,0) model with $\phi_{k0} = 0$, k = 1,2,3, namely

$$F(y_t \mid H_{t-1})$$
$$= .3858\Phi\left(\frac{y_t - .4252y_{t-1}}{.0204}\right)$$
$$+ .5938\Phi\left(\frac{y_t - .1417y_{t-1}}{.0104}\right)$$
$$+ .0204\Phi\left(\frac{y_t}{.1948}\right).$$

The value of BIC is –1704.41. The standard errors of $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\phi}_{11}, \hat{\phi}_{21})$ are (.1630, .1657, .0126, .0034, .0014, .0731, .1176, .0268). This model is similar to the MAR model fitted to the IBM common stock prices data in Wong and Li (2000). Note that that third component models independent pure replacement-type outliers in the return series. Some typical predictive distributions generated from the fitted MAR model are shown in Figure 2. They are either unimodal or bimodal.

For the year 1997, a MAR(3;1,1,1) model with $\phi_{k0} = 0$, k = 1,2,3, is chosen by the minimum BIC criterion. The model is

$$F(y_t \mid H_{t-1})$$
$$= .2473\Phi\left(\frac{y_t - 1.4994y_{t-1}}{.0197}\right)$$
$$+ .6401\Phi\left(\frac{y_t - .1705y_{t-1}}{.0100}\right)$$
$$+ .1126\Phi\left(\frac{y_t - .8713y_{t-1}}{.0244}\right).$$

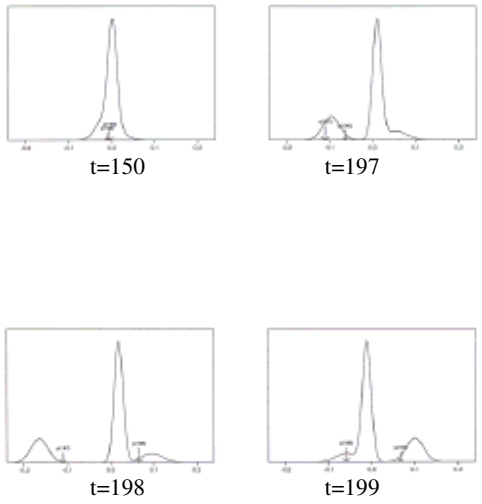

Figure 3: The predictive distribution of the 1997 HSI return series for t = 150, 197, 198 and 199. The actual values at t and t – 1 are also shown.

The value of BIC is –1542.08. The standard errors of $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\phi}_{11}, \hat{\phi}_{21}, \hat{\phi}_{31})$ are (.0553, .0671, .0509, .0026, .0047,1484, .0463, .1235). Here, the estimate of ϕ_{11} is greater than one and hence it is an explosive AR component.

Note that the Hong Kong stock market was greatly affected by the Asian financial crisis and there were many unexpected extreme fluctuations in the market during the

period. Some predictive distributions generated by the fitted MAR model are shown in Figure 3.

During the first half of the year, the stock market is rather peaceful and the predictive distributions are mostly unimodal such as the one for t = 150. However, when the Asian financial crisis began in October, the stock market became extremely volatile. The predictive distributions in the second half of the year are mostly trimodal as those for t = 197, 198 and 199.

We believe that the MAR model will have many potentials in risk management. We hope to see more applications of it in the future.

REFERENCE

Chan, K.S. and Tong, H (1998)
A note on testing for multi-modality with dependent data.
Unpublished manuscript.

C.S. Wong and W.K. Li (2000)
On a mixture autoregressive model.
J. Roy. Statist. Soc. B, 62, pp95-115

C.S. Wong and W.K. Li (2001a)
On a mixture autoregressive conditional heteroscedastic model.
J. Am. Statisti. Ass. (To appear)

C.S. Wong and W.K. Li (2001b)
On a logistic mixture autoregressive model.
Biometrika (To appear)

An outing for Hong Kong Statistical Society

Kelvin Yau
Programme Secretary, Executive Committee

On 7 Jan 2001, we joined the HKSS one-day tour activity. There were 20 members (together with their family members) participating this activity, including those from the Census and Statistics Department (Alvin Li, CM Luk, KC Leung, Karen Chan and Frank Fong), universities (WK Li, Q Fan, TS Lau, MG Gu and JL Tang) and private sector (Lucy Kwan).

At 8:30am, we gathered at Kowloon Tong MTR station. We visited three sight-seeing locations. After making wishes at the "Wishing Tree" in Tai Po, we went to the "Dragon Hill Temple" and the "Greenfields" near the Sai Kung area.

An instructor guided us throughout the "Greenfields" journey, introducing us some interesting things about various types of plants and food chains. Some of us bought a few small plants and fresh vegetables.

Thereafter, we had a buffet lunch in the Newton Hotel. During the lunch, Alvin, Lucy and others won prizes in the lucky draw. The tour ended at around 2:30pm.

News

A Warm Congratulation to Professor Howell Tong!!!!

Professor Howell Tong of the University of Hong Kong has recently been awarded the National Natural Science Prize (China) for his important research in non-linear time series analysis. Being the *only* winner from Hong Kong across all disciplines and the *only* mathematician winning the award in 2000, Professor Tong's entry was based on his two books: Threshold Models in Non-linear Time Series Analysis (Springer-Verlag, 1983) and Non-linear Time Series: a dynamical system approach (Oxford University Press, 1990).

Mrs. Fanny Law, the Secretary for Education and Manpower of the HKSAR government, in her congratulation letter to Professor Tong, said that Professor Tong had set a good example with his outstanding achievement for others to follow and provided an incentive for the higher education sector in Hong Kong.

Earlier in February, the prize was presented at a ceremony in Beijing attended by major state leaders.