Challenges to Time Series Analysis in the Computer Age

By Clifford Lam
London School of Economics and Political Science

A time series course is usually included as a "complete package" to an undergraduate degree in statistics. Despite the fact that there are abundant applications in economics/finance which you may have/have not seen, and that there are many books having both "time series" and "econometrics" in their titles, the use of time series analysis is by no means restricted to these fields. In this article, we briefly explore the objectives and methods of time series analysis in some scientific fields. At the same time, we get a sense of how the enormous computing power of modern days computers has helped statisticians tremendously in their ever-challenging tasks in time series analysis.

Time series and Time Series Analysis

Time series is simply a set of observations recorded at different time points. Since time runs forward, time series observations has a natural ordering. This distinguishes time series data from others, since data points "close" in time are envisaged to share more common characteristics than those which are further apart in time. One common

measure of such characteristics is the correlation of data at different time points, which in time series terminology is called "autocorrelation".

Time series analysis, as the name suggests, is the analysis of time series. But what is meant by "analysis"? Analysis of time series is in particular the study of the autocorrelations in the data. In fact, such information from the data is rich, and can help us understand a lot about the time series at hand. Hence many of the classical or more recent time series analysis methods are set to understand, or "model" these autocorrelations. With these information, we can make forecast for our data, like forecasting earnings next season, or temperatures for next week, etc.

To understand autocorrelations, a time series "model" is usually built, which is used to capture the essential features in the data. If you have studied a time series course, most probably the simplest model you have come across is the AR or MA model. An AR(1) model can be written as

$$y_t = \phi y_{t-1} + \epsilon_t,$$

where y_t is the observed value at time t, ϕ is a real parameter, and ϵ_t represents something we cannot explain - the "noise" at time t. This model implies that the current observed value depends mainly on the immediate past value. And some simple further assumptions and calculations show that the autocorrelation is then decaying exponentially as time gap increases. The principle of time series analysis is that we "build" a model which is as simple as possible, and matches the autocorrelation information from the data at the same time the best we can.

There are many examples of time series and methods for analysis. We give examples of multivariate and high dimensional time series data, where estimation and forecasting are made possible thanks to modern days computing power.

Climate Change

One global question which concerns everyone of us is that if the global temperature is actually rising over the years. Consensus is gradually reached over the past decade and the answer is, unfortunately, affirmative. You may think that this question should be easy enough to answer - just to see if there is an upward trend in the temperature data! This is correct, if we had the temperature data over the past hundreds or thousands of years. Why do we need so much data? It is because temperature fluctuates wildly. It can be day by day fluctuations, month by month, year by year, or even at a much longer time scale. With all these fluctuations, we need to have a long time series of temperature data to confirm a global warming trend which may appear only within the past 100 years. Otherwise, an upward trend we observe today may in fact be some fluctuations with say 30000 years period, which is well observed and similar to today's trend around 30000 years ago.

We do not have such a long time series at hand, however. Hence proxies are measured to "approximate" the climate in the past. Measurements serving as proxies include tree rings, ice core records and sediments under the sea etc. These proxies give us an idea of a general climate of the earth in the past, but not necessarily the accurate measurement of temperature. Hence on top of temperature data, we also study other time series data such as atmospheric pressure, or sea level pressure, which is closely related to temperature.

We introduce a set of MSLP data which is measured over a large region over the North Atlantic Ocean (data obtained from the British Atmospheric Data Centre). The region is divided into grid points according to locations (measured by longitude and latitude). At each grid point, daily sea level pressure is measured, and mean taken

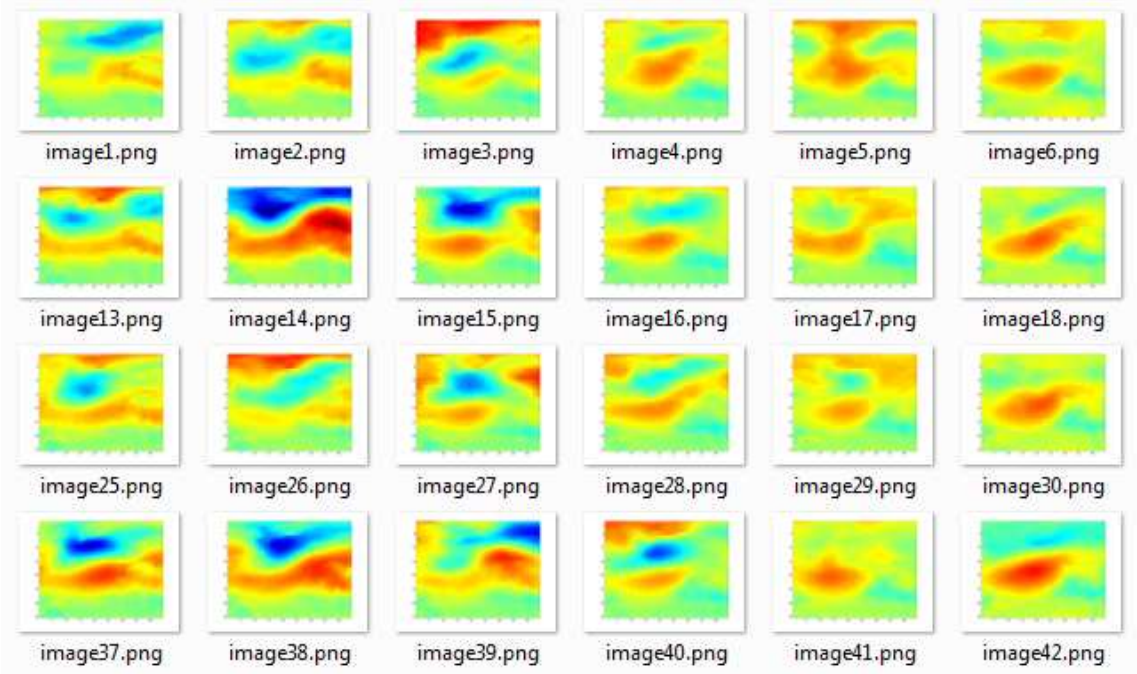

Figure 1: *Mean sea level pressure (MSLP) data over a region around North Atlantic Ocean. Image 1 represents January 1958, Image 6 represents June 1958, Image 13 represents January 1959, and so on. Different intensities on the plots represent different pressure levels.*

over a month period to produce the MSLP data. There are 1540 grid points on all of the images in Figure 1. Since at each grid point we observe a time series of MSLP data (from January 1958 to December 2001), there are 1540 time series each with length $(2001 - 1958 + 1) \times 12 = 528$.

From figure 1, it is clear that there is a strong yearly seasonality in the data. Also, winter (first two columns) and summer (last two columns) data look very different. Are there any other underlying seasonality that can also be discovered from the data? In this region of the world, there is a famous climate phenomenon called the "North

Atlantic Oscillation". Can we recover this from the data? How can we forecast sea level pressure for the region, and see if there are recent "trends" that are not observed in the past? Answering these questions is important to partially understand climate change in the region and the world. Proper time series analysis gives great insights into understanding the answers to these questions.

In China, there are a lot of important problems related to the global climate change. One important example is the monitoring of agricultural activities. Traditional seasonal cycles used by farmers in China are slowly drifting away from the true climate because of global warming. Without adjustments of traditional agricultural activities, food production can be seriously hindered. With a population officially over 1.3 billion, this can have serious consequences in China. Understanding climate change in China and the world is the first step to making suitable national adjustments, while a grip of regional climate change can aid local farmers in their adjustment in agricultural activities.

In the face of modern computers, you may think that the MSLP data is not a large data set. However, it depends solely on what models or methods you are considering for analysis. For instance, if you are considering the vector autoregressive (VAR) model of order d,

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_d\mathbf{y}_{t-d} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{y}_t$ is a vector of 1540 components, then each coefficient matrix $\mathbf{A}_j$ is of size 1540×1540! If $d = 10$, then the total number of parameters needed to be estimated is $1540^2 \times 10 = 23.7$ million! There are just too many parameters relative to the number of data points we have. Hence sometimes researchers will restrict a large number of those parameters to be zero, believing that many of the parameters in the matrices $\mathbf{A}_j$ are 0 or very close to 0. This leads to a huge restricted least square problem for

estimating the $\mathbf{A}_j$'s, which we rely very much on a fast computer and an efficient algorithm to solve.

Needless to say, when we look at other climate variables together like temperature, air pressure, rainfall, etc, analysis becomes much more complicated. How do all these climate variables affect the economic growth globally? How climate change affects the food production and GDP growth in China? The sheer size of data available to us make all these questions even more difficult to answer.

Speech Recognition

Speech recognition should not be new to many of you. Your iphone/BlackBerry may already have some advanced voice recognition tools built in. But you may wonder what have been processed in order that your phone can answer you back? You can be assured that MUCH information have been processed before your phone can answer your question!

In figure 2, a time series representation of my surname is presented. The Matlab function `wavrecord` is used to record what I have pronounced, changing sound wave into digital time series data. An obvious question is, can we recognize from this waveform the original word that is said? If we consider all languages in the world, then it will be an extremely difficult task since many words pronounced similarly in many different languages. However, if we only consider one language, then the task is "much easier". I quoted "much easier", since easy or not still largely depends on what task we want to perform. For instance, if instead of one word, we want to recognize a phrase, or even a sentence?

Figure 3 is the time series representation of my full name. Obviously when we say a

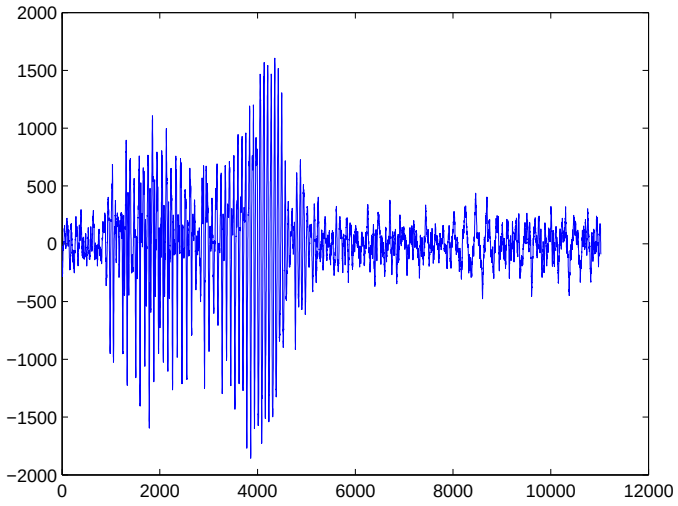

Figure 2:

The Chinese word "林", pronounced in Mandarin. The time interval is 1 second.

phrase or someone's full name, we do not utter word by word, but say it as a complete sentence. Hence my name has three words, but the time series representation in figure 3 shows as if there are only two words pronounced (only two parts of the data are obviously separated from each other). Hence, even if a well developed algorithm is able to detect which Chinese word does the time series in figure 2 represent, it may still find it difficult to recognize the whole phrase in figure 3. On top of this difficulty, notice that between words and after words in both figures 2 and 3, the wave form is not flat, but fluctuates. It is because, obviously, we live in a noisy environment, and even if you pronounce the same words/phrase, background noise can be different. Therefore our algorithm has to distinguish between signals and the background noise at possibly different background noise levels. For more advanced readers, this preliminary step belongs to "denoising" of signal, and is under lots of researches even nowadays. A

common approach for speech signal denoising involves wavelet modeling of the speech time series.

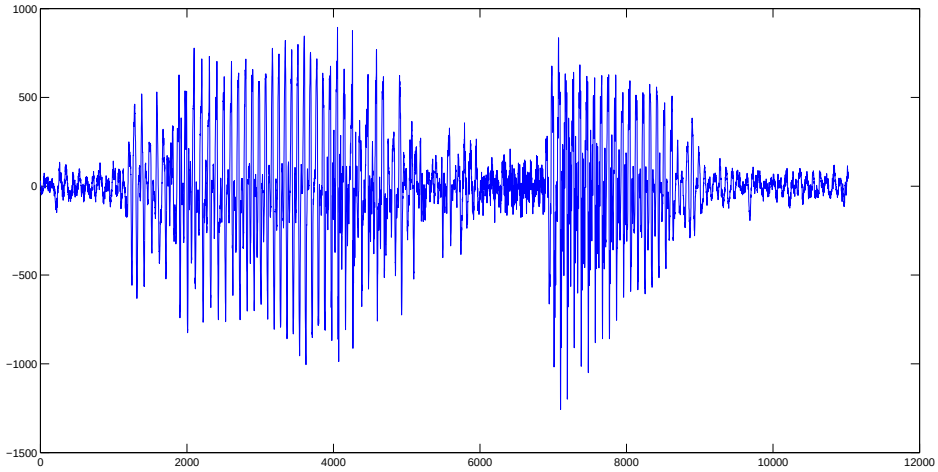

Figure 3:

The Chinese words "林偉峰", pronounced in Mandarin. The time interval is 1 second.

Suppose the signal is now denoised. How to proceed with signal detection? There are two cases. One case is to identify speeches potentially from different speakers, e.g. a public enquiry machine in a library. A general idea is to use a lot of signal data, ideally including all possible words and phrases to be recognized said from a lot of individuals, to "train" up a statistical model. More individuals means that the model can capture more variations of the same words/phrases, increasing accuracy of detection. Machine can be trained and be more accurate in speech detection as more people asked questions, thus more data for the underlying model.

For your phone/computer, it maybe the case that the speech detection tool is set to detect YOUR speech only. It is then an "easier" task, where by talking to

the computer/phone more times, it has more data stored and thus the model built inside is trained better. It is easier than the above task since your own voice has less variations when saying the same words/phrases. Methodologies of speech detection include the very common Hidden Markov Model (HMM), or Neural Network (NN), or a combination of both. This is a very broad area of heated researches nowadays. So don't be surprised if the next generations of computers/mobiles have a much stronger speech detection ability!

Evolution of Networks

When we talk about network, one may immediately think about social networks like Facebook, Twitter, or in China, QQ. These form large networks of people who connected as friends on the internet. Networks of course appear in many scientific fields. Even when you commute from one place to another, you are traveling in a network: bus networks, train networks, or even cycling networks. Within your body, a lot of proteins are linked to same category of body functions, and different group of proteins together may perform more complex functions. Hence looking at a hierarchy of body functions, proteins can be linked together by network. So, what exactly is a network?

A network is a graph with nodes and edges linking them. Individuals with certain characteristics that can be linked together form a network. In all the social networks, pair of friends forms a link between two individuals, and people linked together this way form a network. Buses or train traveling from one place to another form a link between two places, and hence places are linked together by bus or train routes as a network. Proteins or genes are linked together in a network by looking at their functions.

Consider a very simple example, where you record the expenditures of yourself (U), your parents (P), and your sister (S) each day over the last year. Suppose you and your sister's income sources are solely from your parents' weekly pocket money, and you and your sister are not saving up consistently (i.e. in a day you may save up, in another day you may use up some of the savings). Furthermore, suppose unfortunately that you and your sister are not "friend" with each other. What do you think a graph connecting the three variables U, P and S will look like?

When your parents' expenditure is large, they are more likely to give you and your sister less to spend. So in a sense you and your sister's expenditure will be related through how much your parents give to you. In a period where you can spend more, your sister may probably spend more also because your parents give both of you more money, and vice versa. Hence you expect that U and S are actually correlated. However, if given how much money your parents have spent, then you will know how much money you and your sister have approximately. Since you are your sister are not "friends", you two are spending independently on different things given your weekly budget. Hence given P, you expect that U and S are actually independent. In statistics, we say that U and S are **conditionally independent** given P. A graph connecting these three variables should be like that on the left panel of figure 4. A missing link between U and S shows that U and S are not "linked" given P.

After some time, you and your sister fortunately become "friends" with each other! You two spend more time shopping or buy things together, and even share money in buying better stuffs. What will the graph become then? In this case, even given how much your parents' pocket money is for the week, you and your sister's expenditure will be much more alike since you go shopping together and share some money to buy things. On the right panel of figure 4 is a graph showing the new relationship. Note

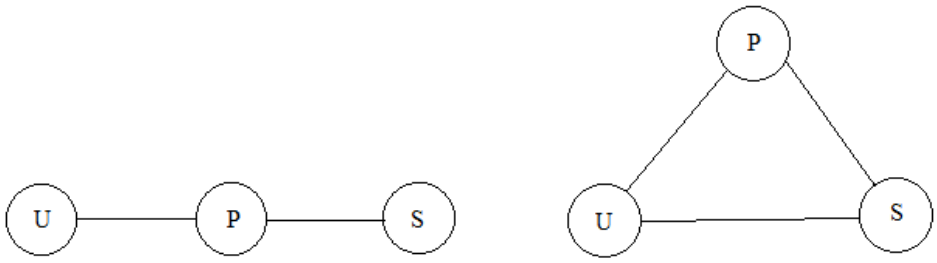

Figure 4: *Left: The graphical model for the variables U, P and S when you and your sister are not "friends". Right: Same, but you and your sister becomes "friends".*

that U and S are now dependent even given P. So from the left panel of figure 4 to the right panel, it tells you how the network of the three variables U, P and S is changed, when you and your sister become "friends".

Since you have the daily expenditure data over the past year, you can use the data to verify the graphs. But how? If you calculate the covariance matrix $\mathbf{\Sigma} = \text{var}(\mathbf{E})$ of the vector $\mathbf{E} = (U, P, S)^T$, and calculate the inverse of this 3 by 3 matrix $\mathbf{\Omega}$, then since U is conditionally independent of S given P, $\mathbf{\Omega}$ should have the form

$$\mathbf{\Omega} = \begin{pmatrix} \bullet & \bullet & 0 \\ \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \end{pmatrix}.$$

So when you use your data before you "befriend" your sister to estimate the inverse of the covariance matrix for the expenditures, or otherwise called the **precision matrix** or **concentration matrix**, the $(3, 1)$th entry should be small and close to zero. And after you and your sister are friends again, the estimated inverse covariance matrix should not have close-to-zero entries.

The above example gives you an idea of how some graphical models are estimated. The problem can be translated to estimation of a **sparse** precision matrix, where a zero

entry in the precision matrix means that the corresponding variables are conditionally independent, given all other variables. In finance, estimating the precision matrix of a set of stock returns is a fundamental tool in asset allocation. Assets (stocks) can then be linked together in a graph, aiding our understanding of how stock returns are related to one another.

As in figure 4, the network is actually changing over time. How can we model or capture the changes? This is important since a change in the network means a change in the optimal allocation in assets. In a social network with millions of users as nodes, modeling the changes of it over time is a key in strategic decisions related to the network. A good model can forecast how the network will evolve over time, and hence tells the business decision maker if something has to be done, like introducing new games, introducing new functions, etc to control the growth. In this case, even a simple model will be difficult to analyze, since the number of nodes is huge, and the number of edges is even bigger. A network with millions of nodes can potentially have trillion edges! Summary and presentation of these huge networks, let alone analysis, present tremendous challenges. Modern computing power has definitely helped, but explosion of data can easily overcome such power. Efficient modeling and analysis are still the keys to success.

Another type of problem is the so called change point detection problem. If your parents suspect that you and your sister have become "friends" again, how can they detect this from analyzing the expenditures data? How can they know approximately when you and your sisters are friend again? How can a hedge fund manager detect a change in the network of stock returns, and hence make appropriate adjustments to asset allocations? Once again, enormous computing power plays an important role in solving these problems.

Of course not all network data can be handled using precision matrix estimation. For instance, if some nodes are linked, but with directional edges - like wind or ocean currents that are moving in one direction only for some time of a year. These data requires more modern treatments of network data, and this area of statistics has attracted more and more attentions from researchers.

Green Cities

As I'm writing this article, one of my MSc student is working on her master dissertation on comparing two "green cities" in Europe: Stockholm and Copenhagen. In the London School of Economics where I'm a lecturer, there is an international centre called "LSE Cities". It studies "how people and cities interact in a rapidly urbanizing world, focusing on how the design of cities impacts on society, culture and the environment", and aims to "shape new thinking and practice on how to make cities fairer and more sustainable for the next generation of urban dwellers, who will make up some 70 per cent of the global population by 2050" (see the LSE Cities web page `http://www2.lse.ac.uk/LSECities/home.aspx`).

Some of the data I got about Stockholm can be downloaded from the web page `http://uskab.se/index.php/statistics-in-english.html`. You better understand a little bit of Swedish to get what the data is exactly about! There are all sorts of economic indicators and environmental variables we need to look at. For example, GDP of the city or aggregate income can both be measures of the economy in a city. These are closely related to the population, the labour market, house prices, educational background of people, unemployment rate, migration rate, etc. The list can go on. And environmental variables? Even more! Air quality (measured by various greenhouse

gas emission from various sources (long list omitted)), water quality (bathing water, lake, bays, etc), sewage components, toxic pollutants in bottom of sediments, contaminations in fish, number of cars and types, driving distances (gasoline consumption), different types of traffic measurements, etc.

There are a number of aims in studying the data. One important objective is to see if "green policies" can have positive impacts on the economy. Stockholm has long been implementing a lot of green policies, like controlling greenhouse gas emissions. It is important to gauge the performance of these policies, and measuring its economic impacts is one way of doing this. Also, for other cities to follow suit, a demonstrated positive impact on the economy or other attributes of a city from these green policies is a key.

So, how can we utilize the vast amount of data to analyze the economic impact of green policies? All variables are time series, but they are not recorded with the same frequency - some are monthly data, some are quarterly and some are just yearly! Majority of them do not go back far in time as well. Even if we just want to use a multiple linear regression, the different recording frequency in the data posts a challenge already. And, even if this can be overcome, there are just too many variables for a linear regression with such short time series. We have not even talked about if a linear regression is appropriate in the first place!

There are many dimension reduction techniques for time series data. Some aims to find a lower dimensional dependence structure for the data, and some focus on finding common dynamics. This is an important first step towards analyzing complex relationships between time series. Innovated algorithms are created for these techniques, and modern computing power again plays an important part in such high dimensional data analysis.

After seeing all these examples, I hope you get a sense of the importance of time series analysis, and how easily modern data becomes huge and difficult to analyze. Ever-increasing computing power plays a crucial role to analysis of large data sets, yet understanding the underlying structure of the data, extracting useful information and communicating them to a broader audience still relies on innovations of statisticians. Every scientific field has their own need of statistics whenever they need to handle data, and many can be time series. Are you up to the challenge?