Chapter 2
Applied Time Series Analysis

Abstract There are many reasons to analyze the time series data, for example, to understand the underlying generating mechanism better, to achieve optimal control of the system, or to obtain better forecasting of future values. Applied time series analysis consists of empirical models for analyzing time series in order to extract meaningful statistics and other properties of the time series data. Time series models have various forms and represent different stochastic processes. Time series analysis model is usually classified as either time domain model or frequency domain model. Time domain models include the auto-correlation and cross-correlation analysis. In a time domain model, mathematical functions are usually used to study the data with respect to time. The three broad classes for modeling the variations of time series process are the autoregressive models, the integrated models, and the moving average models. The autoregressive integrated moving average models are the general class of these models for forecasting a time series that can be stationarized by transformations such as differencing. In a frequency domain model, the analysis of mathematical functions or signals is conducted with respect to frequency rather than time. Mathematical models can be used to convert the time series data between the time and frequency domains. The parameters and features in the frequency domain can be used as inputs for the mathematical models like discrimination analysis and improved results can be obtained.

Keywords Time series analysis • Basic characteristics • Estimation of correlation • Autoregression • Mathematical models
2.1 Basic Characteristics of Time Series

2.1.1 Estimation of Correlation

2.1.1.1 Auto-Correlation Analysis

The autocorrelation of a time series measures the correlation between values of the time series at different points in time, as a function of the two times. It can be used to find repeating patterns, or to identify the missing fundamental frequency in a signal implied by its harmonic frequencies. It is a popular tool for analyzing time domain signals. Mathematically, the autocorrelation between any two time \( s \) and \( t \) is defined by

\[
R(s,t) = \frac{E[(X_i - \mu)(X_s - \mu)\bar{X}_s]}{\sigma_i \sigma_s}
\]

where \( X_i \) is the value of a given time series \( X \) at time \( i \), \( \mu \) and \( \sigma_i \) are the mean and variance for all time \( i \). If the mean and variance of the time series are time-independent, the autocorrelation can be simplified as

\[
R(\tau) = \frac{E[(X_i - \mu)(X_{i+\tau} - \mu)\bar{X}_s]}{\sigma^2}
\]

where the time-lag \( \tau = t - s \). For a discrete process of length \( n \{X_1, X_2, ..., X_n\} \) with known mean and variance, an unbiased estimate of the autocorrelation is

\[
\hat{R}(k) = \frac{1}{(n-k)\sigma^2} \sum_{i=1}^{n-k} [X_i - \mu][X_{i+k} - \mu]
\]

for any positive integer \( k < n \). If the mean and variance are replaced by the standard formulae for sample mean and sample variance, then it is a biased estimate (Priestley 1982). A fundamental property of the autocorrelation is symmetry, i.e. \( R(i) = R(-i) \). Another property is that the continuous autocorrelation function reaches its peak at the origin. It can also be shown that the autocorrelation of a periodic function is also periodic with the same period.

2.1.1.2 Cross-Correlation Analysis

Cross-correlation is a measure of similarity of two time series as a function of a time-lag applied to one of them. It can be applied to search a long duration signal for a shorter, known feature, and has applications in areas like pattern recognition and single particle analysis, etc. Mathematically, the cross-correlation is defined as

\[
(f \ast g)(t) = \int_{-\infty}^{\infty} f^*(\tau)g(t + \tau)d\tau
\]
for continuous time series functions, \( f \) and \( g \), where \( f^* \) is the complex conjugate of \( f \). For discrete functions, the cross-correlation is defined as

\[
(f \cdot g)[n] = \sum_{m=-\infty}^{\infty} f^*[m] g[n+m]
\]

For two independent random variables \( X \) and \( Y \) with probability distributions \( f \) and \( g \), the probability distribution of the difference \( Y-X \) is given by the cross-correlation \( f \cdot g \). A property of the cross-correlation of functions \( f(t) \) and \( g(t) \) is \((f \cdot g) \cdot (f \cdot g) = (f \cdot f) \cdot (g \cdot g)\). The cross-correlation of a time series with itself is the autocorrelation, and there will always be a peak at a lag of zero. In template matching, the normalized correlation can be defined as (Lewis 1995):

\[
\gamma(u, v) = \frac{\sum_{s,t} [f(s,t) - \bar{f}_{u,v}][t(x-s, y-t)-\bar{T}]}{\left\{ \sum_{s,t} [f(s,t) - \bar{f}_{u,v}]^2 \sum_{s,t} [t(x-s, y-t)-\bar{T}]^2 \right\}^{0.5}}
\]

where \( \bar{T} \) is the mean of the feature and \( \bar{f}_{u,v} \) is mean of \( f(x,y) \) in the region under the feature.

### 2.1.1.3 Autocorrelation Functions

The autocorrelation function can identify the time interval over which a correlation in the noise exists. As said in the first section, autocorrelation is the correlation of a time series set with itself, offset by \( n \)-values. The lag \( k \) autocorrelation function is the set of autocorrelations with offsets 1, 2, 3, 4,\ldots, \( k \). Mathematically, the lag \( k \) autocorrelation function is defined as

\[
\tau_k = \frac{\sum_{i=1}^{N-k} (f(i) - \bar{f})(f(i+k) - \bar{f})}{\sum_{i=1}^{N} (f(i) - \bar{f})^2}
\]

where the observations are equi-spaced.

The autocorrelation function can be used to check the non-whiteness of data or residuals, to detect periodic components in data, and to identify the dominant power law noise type. The autocorrelation function of a periodic signal is itself a periodic signal, with a period the same as that of the original signal. Short signals can have short autocorrelation functions as well, while the autocorrelation functions of random noise is a single sharp spike at shift zero. It is interesting to note that the autocorrelation function and the power spectrum are related by the Fourier transform (Percival and Walden 1993). When the Fourier transform is applied to the autocorrelation function, the power spectrum can be obtained.
2.1.2 Stationary Time Series

A time series is said to be stationary if its statistical properties do not depend on time. A time series may be stationary with respect to one characteristic, while not stationary with respect to another. A time series is said to be weakly stationary if its mean, variance and autocovariance do not grow over time. For example, there can be a time series where its mean does not depend on time while its variance depends on time. As an example, white noise is stationary. A time series is said to be a white noise process if each value in the time series have zero-mean, constant conditional variance and is uncorrelated with all other realizations. Another example is the autoregressive moving average model, which is discrete-time stationary process with continuous sample space. Many business time series are far from stationary. Nevertheless, stationary time series can provide us with meaningful sample statistics such as means, variances, and correlations with other variables. Also many statistical forecasting methods are based on the assumption that the time series are approximately stationary. Mathematical transformations like deflation, seasonal adjustment, and de-trending can be applied to try to stationarize the time series.

2.1.3 Smoothing of the Time Series

The irregularities and random fluctuations in time series data often obscure the efforts to understand the process being observed. Smoothing techniques are employed to address these problems and to make long term fluctuations in the time series stand out more clearly. The moving average smoothing is the most common type of smoothing techniques. The technique can reduce random fluctuations and make adjustments for seasonal or cyclical components of a time series. Mathematically, the moving average smoothed statistic $s_t$ is the mean of the last $k$ observations:

$$s_t = \frac{1}{k} \sum_{n=0}^{k-1} x_{t-n}$$

where the choice of an integer $k > 1$ is arbitrary. A slightly modified version of the simple moving average technique is the weight moving average, which is defined as

$$s_t = \sum_{n=1}^{k} w_n x_{t+1-n}$$

where the set of weighting factors satisfy the constraint $\sum_{n=1}^{k} w_n = 1$. 

Another popular smoothing technique is the exponential smoothing. It can reduce the effects of random fluctuations in the time series and in many cases is an effective tool in time series forecasting. In moving average smoothing, all the past observations are weighted equally, whereas in exponential decreasing weights are assigned as the observations get older. Mathematically, the exponential smoothing can be computed with the algorithm:

\[ s_t = \alpha x_t + (1 - \alpha)s_{t-1} \]

where \( s_0 = x_0 \), and \( \alpha \) is the smoothing factor with the constraint \( 0 < \alpha < 1 \).

### 2.1.4 Periodogram Analysis

The periodogram is an estimate of the spectral density of a signal (Schuster 1898). Mathematically, for an ordered set of \( N \) random variables \( X_1, X_2, \ldots, X_N \), the periodogram can be computed as followed (Herbst 1963):

\[ I_x(\lambda) = 2\left| J_x(\lambda) \right|^2 \quad (0 \leq \lambda \leq \frac{1}{2}) \]

where \( I_x(\lambda) \) is computed at a finite set of \( \lambda \)-values, usually

\[ \lambda = \frac{s}{N} \left( s = 0, 1, 2, \ldots, \left\lfloor \frac{1}{2} N \right\rfloor \right) \]

\[ J_x(\lambda) = N^{-\frac{1}{2}} \sum_{t=1}^{N} X_t z^{\lambda t} = N^{-\frac{1}{2}} \sum_{t=1}^{N} X_t \cos \frac{2\pi \lambda}{N} + i N^{-\frac{1}{2}} \sum_{t=1}^{N} X_t \sin \frac{2\pi \lambda}{N} \]

with \( z^{\lambda t} = e^{2\pi i \lambda t} \) for real finite \( \lambda \), and, for real random variables \( U, V \),

\[ |U + iV|^2 = U^2 + V^2. \]

In raw periodogram, it may not be a good spectral estimate because the variance at a given frequency does not decreased as the number of samples used in the computation increases. The smoothing of the periodogram can be used to address this variance problem. When smoothing is applied to reduce the effect of measurement noise, the smoothed version of the periodogram is also called the spectral plot. Another technique for this problem is the method of average periodograms (Engelberg 2008). In practice, the periodogram is often computed from a finite-length time series using the fast Fourier transform (FFT).
2.2  Autoregression and ARIMA Models

2.2.1 Time Series Regression

Linear regression is useful for exploring the relationship of an independent variable to a dependent variable when the relationship is linear (O’Sullivan and Rassel 1999). However, there are problems when there is the autocorrelation in a time series. In the time series, the values at one time may be influenced by values at a previous time. This happens when the values of the dependent variable over time are not randomly distributed. Another problem is that linear regression fails to capture seasonal, cyclical, and counter-cyclical trends in time series. Regression options like transforming variables with logarithms, differences, inverses, lags, and de-trending can be applied for the regression analysis. The form in the regression relies ultimately upon the time series process of the variables concerned. The theory can serve as a guide to the variable selection, while the regression results may lead to ways to refine the underlying theory.

Time series regression models are very suitable for evaluating short-term effects of time-varying exposures. Nevertheless, there are problems that may occur in the regression models, like multicollinearity, heteroscedasticity, and autocorrelation (Ostrom 1990). The Durbin-Watson statistic can be employed to test for the presence of first-order autocorrelation in the residual of a regression model. Treatments for autocorrelated error terms include the finding of an important omitted variable, transforming of the variables based upon generalized least squares, and introducing of the time as a dependent variable on the regression model.

A popular statistical package for time series analysis is the RATS, an abbreviation of Regression Analysis of Time Series (http://www.estima.com). It is relatively inexpensive as compared with other similar commercial packages. The package includes the major procedures in econometrics and time series analysis, like linear regression, regressions with heteroscedasticity and serial-correlation correction, non-linear least squares, generalized method of moments, maximum likelihood estimation, ARIMA, spectral analysis, Kalman filter, logistic regression, ARCH and GARCH models, and vector autoregressions. Another popular statistical computing package is the R software environment. Its source code is freely available under the GNU General Public License, and pre-compiled binary versions are available for operating systems (http://www.r-project.org). R supports statistical computing like linear and nonlinear modeling, classical statistical tests, time series analysis, classification and clustering, etc.

2.2.2 Autoregressive Moving Average Models

The autoregressive moving average model consists of the autoregressive (AR) and the moving average (MA) parts, and is referred as the ARMA($p,q$) model,
where \( p \) is the order of the autoregressive part and \( q \) is the order of the moving average part (Box et al. 1994). Mathematically, the AR(\( p \)) model can be expressed as

\[
X_t = c + \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t
\]

where \( c \) is a constant, \( \phi_1, \ldots, \phi_p \) are the parameters of the model, and \( \varepsilon_t \) is the white noise. In the AR(1) model, the model is not stationary for \( |\phi_1| \geq 1 \). Mathematically, the moving average MA(\( q \)) model can be expressed:

\[
X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}
\]

where \( \mu \) is the expectation of \( X_t \), \( \theta_1, \ldots, \theta_q \) are the parameters of the model, and the \( \varepsilon_t, \varepsilon_{t-1}, \ldots \) are the white noise error terms. Combining the AR(\( p \)) and MA(\( q \)) models, we can have the mathematical expression of the ARMA(\( p, q \)) model:

\[
X_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_i \varepsilon_{t-i}
\]

In the above models, the error terms are assumed to be independent identically-distributed random variables (i.i.d.) sampled from a normal distribution with zero mean: \( \varepsilon_t \sim N(0, \sigma^2) \), where \( \sigma^2 \) represents the variance. These ARMA models are univariate models, and extensions for the multivariable cases are the vector autoregression (VAR), which we will talk about in later chapter.

### 2.2.3 Building ARIMA Models

ARIMA (autoregressive integrated moving average) models are the general class of models for forecasting a time series that can be stationarized by transformations such as differencing. The first step in the ARIMA procedure is to ensure the series is stationary (Kosuke et al. 2008). The additional differences are applied until the diagnostics indicate stationarity. This is to derive the stationary time series. The lagged values and model errors of the stationary time series are used to model the time series. The fine-tuning of the model is accomplished by adding lags of the differenced series and/or lags of the forecast errors to the prediction, as necessary so that the autocorrelation from the forecasting errors can be removed as much as possible. ARIMA utilizes a state space representation of the ARIMA model to assemble the likelihood and then utilizes the maximum likelihood to estimate the parameters of the model (Brockwell and Davis 1991).

ARIMA models can be expressed as ARIMA(\( p,d,q \)), where \( p \) is the number of autoregressive terms, \( d \) is the number of non-seasonal differences, and \( q \) is the number of lagged forecast error in the prediction equation. ARIMA(0,1,0) is the random walk
process, while ARIMA(1,1,0) is the differenced first-order autoregressive model. ARIMA(0,1,1) is the simple exponential smoothing model, while ARIMA(0,2,1) or ARIMA(0,2,2) is the linear exponential smoothing model. Mathematically, the ARIMA(p,d,q) model of the time series \{x_t, x_{t-1}, \ldots\} is defined as

\[ \Phi_p(B) \Delta^d x_t = \Theta_q(B) \epsilon_t \]

where \( B \) is the backward shift operator, \( Bx_{t-1} = x_t \), \( \Delta = 1 - B \) is the backward difference, and \( \Phi_p \) and \( \Theta_q \) are polynomials of order \( p \) and \( q \) respectively. It can be observed that the model consists of three different parts, i.e. an autoregressive part \( \Phi_p = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \), an integrating part \( \Delta = B^d \), and a moving average \( \Theta_q = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \).

### 2.2.4 Forecasting and Evaluation

The above models can help us to better understand the time series data, and to forecast future points in the time series. As these models with lagged values are often used for forecasting, researchers have looked for measures that can produce more accurate results for out-of-sample forecasting (Greene 2000). Instead of the adjusted \( R^2 \) (sum of squares of the model errors), measures like Akaike information criterion (AIC), and Schwartz Bayesian criterion (SBC) are utilized, with

\[ \text{AIC} = T \text{ln}(\text{residual sum of squares}) + 2n \]

\[ \text{SBC} = T \text{ln}(\text{residual sum of squares}) + n \text{ln}(T) \]

where \( n \) is the number of parameters estimated (\( p + q + \) possible constant term), and \( T \) is the number of usable observations. The value of \( n \) yielding the minimum AIC value specifies the best model, and AIC tends to be superior in smaller samples where the relative number of parameters is large (McQuarrie and Tsai 1998). The idea behind these measures is that incorporating additional coefficients will necessarily increase fit at a cost of reducing degrees of freedom (Enders 1995). The objective is to have parsimonious model that fits the data well without incorporating any needless coefficients. The parsimonious models can generally produce better forecasts than over-parameterized models.

### 2.2.5 Causality of the Time Series

The concept of the causality of the time series was introduced by Wiener (1956) and Granger (1969) and become a popular notion for studying the dynamic relationships between time series. A time series variable \( Y \) is said to cause \( X \) in the
2.3 Mathematical Models in the Frequency Domain

2.3.1 Introduction

In the frequency domain, the analysis of mathematical functions or signals is conducted with respect to frequency rather than time. Mathematical models can be used to convert the time series data between the time and frequency domains. A time series can be decomposed into cyclic components, which can be described by their periods and frequencies. The frequency of a cycle is the number of occurrences...
in a fixed time unit. The analysis in the frequency domain can detect these cycles and compute their corresponding frequencies. The Fourier transform is a popular transforming tool to compute the frequency domain representation of a time series. It can generate a description of the distribution of the energy (the amplitude) of the time series as a function of frequency. The plot of this amplitude (y-axis) against the frequency (x-axis) is called the spectrum. In practice, the Fast Fourier Transform (FFT) is often used as a quick way of performing this transform. The parameters and features in the frequency domain can be used as inputs for the mathematical models like discrimination analysis and improved results can be obtained (Voss et al. 1996; Jing and Zhang 2004).

2.3.2 Discrimination Analysis

Discrimination analysis is a method for classifying a set of observations into predefined classes, based on a set of input variables known as predictors. It is very useful for the detection of variables that can discriminate between different groups efficiently, and for the classification of new cases into different groups (Hill and Lewicki 2007). The model is built from a training set of observations with known classes. Computationally, discrimination analysis is very similar with the analysis of variance (ANOVA). Its underlying idea is to decide whether groups differ with regard to the mean of a variable, and then to utilize that variable for the prediction of group membership of new cases. As a simple example, suppose that there is a random sample of 100 adults and 100 children. On the average, adults are heavier than children. Thus, the variable weight can be utilized to discriminate between adults and children with a better than chance probability. In practice, several variables may be considered in a single study to check which variables contribute to the distribution between groups. Multivariable F tests can be used to test whether there are any significant differences with regard to the variables between groups.

The objective of discrimination analysis is to find the variables that discriminate between groups. There are several popular models to implement the discrimination analysis, including the forward stepwise analysis, backward stepwise analysis, and F to enter, F to remove technique, etc. In forward stepwise analysis, at each step, all the available variables are checked to see which one contributes most to the discrimination between groups. Then that variable will be included in the model, and the testing will proceed to the next step with all the other still available variables. In the backward stepwise analysis, firstly all variables are included in the model. Then, at each step, the variable that contributes least to the predicting of group membership will be eliminated. The procedure will stop when the contribution of each remaining variable to the membership prediction is higher than the predefined threshold value. In the F to enter, F to remove strategy, it is similar with the stepwise multiple regression procedure, and the stepwise procedure follows the F to enter and F to remove values.
The discrimination analysis can be applied in bankruptcy prediction, face recognition, marketing research, and product management, etc. (McLachlan 2004). The analysis can also be applied to the input variables in the frequency domain. For example, Jing and Zhang (2004) proposed a new face and palmprint recognition approach based on the discrete cosine transform (DCT) and discrimination analysis. A two-dimensional separability judgment was used to select the DCT frequency bands with favorable linear separability. Then, the linear discriminative features were extracted from these selected bands for the classification process. The procedure can result in better classification performance, significantly improve the recognition rates and effectively reduce the dimension of feature space.

2.3.3 Clustering Analysis

The goal of the clustering algorithms is to figure out the underlying similarities among a set of feature vectors $x$, and to cluster similar vectors together (Theodoridis and Koutroumbas 2003). The clustering process can also be called unsupervised pattern recognition. This is different from the supervised pattern recognition, in which a set of training data is available, and which the classification algorithms can exploit this known information in advance. The clustering algorithms have many different applications in social sciences, engineering and medical science.

The clustering process can be viewed as a combinatorial problem of putting the data points into optimal clusters. However, it is NP-hard to enumerate all such possibilities of clustering. Let $S(N, m)$ be the number of all possible clustering of $N$ vectors into $m$ groups. We can easily see that $S(N, 1) = 1$, $S(N, N) = 1$, and $S(N, m) = 0$, for $m > N$. It satisfies the following recursive relationship (Spath 1980):

$$S(N, m) = mS(N-1, m) + S(N-1, m-1)$$

Its solution is found to be the Stirling numbers of the second kind (for details, see (Liu 1968), etc.):

$$S(N, m) = \frac{1}{m!} \sum_{i=0}^{m} (-1)^{m-i} \binom{m}{i} i^N$$

We can see that the solutions for this problem explore exponentially with (Spath 1980): $S(15,3) = 2.38 \times 10^6$, $S(20,4) = 4.52 \times 10^{10}$, and $S(100,5)$ is of order $10^{68}$. It is impractical to enumerate all possible clusters for all possible values of $m$.

The clustering algorithms can also allow us to infer some hypothesis concerning the nature of the data. It can be a tool for suggesting hypothesis (Theodoridis and Koutroumbas 2003). These hypotheses can be verified by using other data sets as validation sets. Another use is on the prediction that bases on groups. The algorithms can provide us with clusters that are characterized by the similarity of vectors within each cluster. When a new data set or pattern is available, we can assign
it to the known cluster by comparing its characters with each cluster’s characters. Clustering is important for the data reduction purpose too. There are many times that the amount of the data is very large and it is expensive to process all the data. Cluster analysis can be employed for grouping the data into a number of clusters, and then we can process each cluster as a single element.

In view of the computational difficulty, different clustering algorithms have been developed so that only a small number of the different possible combinations of the clusters will be considered. There are four main types of clustering algorithms: sequential algorithms, hierarchical clustering algorithms, clustering algorithms with cost function optimization, and others (like branch and bound algorithms, and genetic clustering algorithms). Different clustering algorithms usually produce different clustering results. It may depend on the problem to decide which type of clustering algorithms is employed.

In designing the clustering algorithm for solving a problem, a major issue is on how to define the similarity between two feature vectors. It is important to choose an appropriate measure for this task. Then, it is also important to choose an appropriate algorithmic scheme that clusters the vectors, basing on the selected similarity measure. Generally speaking, different results can be obtained with different algorithmic schemes. Expert opinions are often needed for the interpretation of the results and for choosing a suitable scheme. In our project, experts from the Genome Research Centre have provided us expert opinions on this subject.

The concept of distance measure is important for the clustering process, which need this measurement of the mathematical distance between individual observations, and groups of observations (Finch 2005). Distance in this context can be in the Euclidean sense, or some other comparable conceptualization like Manhattan distance, Hamming distance, etc. This will affect the shape of the clusters, as some objects may be close to one another with one distance while further away with another distance. A primary assumption underlying these distance measures is that the variables are continuous in nature. Finch discussed about the distance measures in Cluster analysis with dichotomous data. The definition of a distance measure has an important role in the evaluation of clustering algorithms of gene expression profiles. Ido et al. (2007) compared different clustering solutions when using the Mutual Information (MI) measure, Euclidean distance and Pearson correlation coefficient.

Clustering process can be grouped as hierarchical or partition clustering. Hierarchical clustering find successive clusters based on previously established clusters. Partition clustering establishes all clusters at once. Hierarchical clustering can be further divided into two basic groups: agglomerative and divisive clustering. Agglomerative clustering is of the bottom-up approach, that is, starts with each object as a separate cluster and then merge the objects into successively larger clusters. On the other hand, the divisive clustering is of the top-down approach. It starts with the whole set and then divide it into successively smaller clusters. Co-clustering is a clustering technique that not only clusters the objects themselves, but also their features as well. Another difference among the clustering algorithms is to look at whether the clustering process uses symmetric or asymmetric distances.
For example, Euclidean distances are symmetric, which mean distance from object H to K is the same as the distance from K to H. In applications like sequence-alignment methods, asymmetric distances have been used (for example, Prinzie and Van den Poel 2006).

The partition clustering algorithms usually determine all the clusters at once. K-means clustering, fuzzy c-means clustering and derivatives are popular partition algorithms. In the K-means algorithm, initially there are \( k \) randomly generated clusters. The average of all the points in a cluster is assigned as the center. Then, the data point is assigned to the nearest cluster center, and the new cluster centers are computed. The above two steps are repeat until the convergence criterion is met. In fuzzy clustering, each data point belongs to a cluster to a certain extent, rather than to one certain cluster completely. Besides this difference, the fuzzy c-means algorithm is similar with the k-means algorithm. Like hierarchical clustering, the partition clustering algorithms are also very popular for the genomic analysis, with microarray data sets, etc.

In spectral clustering, the dimensionality reduction for clustering in lower dimensions is performed with the spectrum of the similarity matrix of the data. A popular spectral clustering is the Shi-Malik algorithm, which is widely used for image segmentation. Liu et al. (2008) applied the spectral clustering to the analysis of correlation mutations in HIV-1 protease. The spectral clustering of the resulting covariance matrices disclosed two distinctive clusters of correlated residues. Oliveira and Seok (2008) proposed a multilevel spectral algorithm which can identify protein complexes more accurately with less computational time. Winter et al. (2004) studied the problem of blind source separation (BSS) with complex valued sample data in the frequency domain. High quality separation of speech sources is important for the further processing like speech recognition, and BSS is about the finding of the unknown underlying mixing process. Winter et al. proposed an algorithm based on hierarchical clustering to estimate the mixing matrix. The algorithm directly works on the complex valued sample data in the frequency domain and shows good convergence. Accurate estimations of the mixing matrix and very low musical tone noise were achieved.

### 2.3.4 Principal Components and Factor Analysis

Among the tools of the dimension reduction and transformation, the principal component analysis (PCA) is a popular tool for many researchers. Its basic idea is to find the directions in the multidimensional vector space that contribute most to the variability of the data. The principal component analysis was applied to reduce the dimensionality of the gene expression data in studies (Taylor et al. 2002; Yeung and Ruzzo. 2001, etc.). The focuses are on the effective dimensional reduction by the PCA, the analysis of the compressed space and the assistance of the PCA for the classification and the clustering. Khan et al. (2001) applied the PCA and neural network for the classification of cancers using gene expression profiling.
Its basic idea is to find the directions in the multidimensional vector space that contribute most to the variability of the data. The representation of data by the PCA consists of projecting the data onto the k-dimensional subspace according to

\[ x' = F(x) = A'x \]

where \( x' \) is the vectors in the projected space, \( A' \) is the transformation matrix which is formed by the \( k \) largest eigenvectors of the data matrix, \( x \) is the input data matrix. Let \( \{x_1, x_2, ..., x_n\} \) be the \( n \) samples of the input matrix \( x \). The principal components and the transformation matrix can be obtained by minimizing the following sum of squared error:

\[
J_k(a, x') = \sum_{h=1}^{n} \left( m + \sum_{i=1}^{k} a_{hi} x'_i \right) - x_h^2
\]

where \( m \) is the sample mean, \( x'_i \) the \( i \)th largest eigenvector of the co-variance matrix, and \( a_{hi} \) the projection of \( x_h \) to \( x'_i \).

Principal component analysis can be applied with data in the frequency domain. As an example, it is used as a general exploratory tool which leads to specific generator models in quantitative electroencephalography (qEEG). A great part of qEEG is about the computing and interpreting of EEG spectra in the frequency domain (Valdes et al. 1992). Valdes et al. proposed a new method for estimating sources in the frequency domain which fitted dipoles to the whole crossspectrum. The PCA was used for the exploratory data analysis, and the spherical harmonic functions are used to explain the recurring pattern of maps characteristic of the spatial PCA of qEEG data. Wu et al. (1998) applied the frequency spectrum PCA for modeling the sound frequency distribution features. The frequency spectra of the found were treated as a vector in a high-dimensional frequency feature space. The PCA was used to compute the variance distribution for the frequency vectors, with the largest eigenvalues accounting for the most variance within the data set. The proposed method is shown to be simple and reliable for acoustic identification.

**2.3.5 Dynamic Fourier Analysis**

Dynamic Fourier analysis is useful for the non-stationary time series analysis, by giving a local-time representation of the spectrum (Shumway and Stoffer 2006). In some cases, the focus of the study is about the non-stationarity of a time series, i.e. the researcher may interest in the local behavior of the process, instead of the global behavior of the process. The dynamic model starts with the creation of sub-samples (blocks) of the time series data. Then, the spectral density on each block is computed separately. Finally, the spectral densities are plotted together in a 3D plot. Mathematically, the local Fourier transforms of the time series \( x_i \) can be expressed as
2.3 Mathematical Models in the Frequency Domain

\[
d_{j,k} = n^{-\frac{1}{2}} \sum_{r=1}^{n} x_r \psi_{j,k}(t)
\]

where

\[
\psi_{j,k}(t) = \begin{cases} 
(n/m)^{1/2} h_t e^{-2\pi i j t/m} & t \in [t_k + 1, t_k + m] \\
0 & \text{otherwise}
\end{cases}
\]

where \(h_t\) is a taper, \(m\) is some fraction of \(n\), \(j\) indexes frequency, and \(k\) indexes the time shift of the transform. The transforms here are based on local sinusoids, i.e., the tapered cosines and sines that have been zeroed out over various regions in time.

It is increasingly popular in medicine to use the frequency-domain analysis of short-term heart rate variability (HRV) measurements (Hartikainen et al. 1998). Spangl and Dutter (2007) used overlapping windows to access the heart rate variability in the frequency domain. Robust dynamic Fourier analysis of the original short-term HRV data was conducted. As another application example, the dynamic Fourier analysis can also be used for the time series analysis of the explosion and earthquake series (Shumway and Stoffer 2006). A Fourier analysis is performed on a short section of the time series data. Then, the section is shifted, and a Fourier analysis is performed on the new section, and so on until the end of the time series. In this way, the spectral behavior of the signal evolving over time can be obtained.

2.3.6 Random Coefficient Regression

Random coefficient regression (RCR) is known to be robust against data that are not missing completely at random (Laird and Ware 1982). The model can handle multivariate ridge-type regressions and inversion problems, and the usual random effects analysis of variance in the frequency domain is a special case in the RCR. Different from the regression models discussed in the above sections, the regression vector of RCR is regarded as unknown stochastic signal. There are many cases in time series analysis that this assumption is more natural. Mathematically, a RCR model can be expressed as (Feuerverger and Vardi 2000):

\[
Y_i = A_i + B_i X_i, \ i = 1, \ldots, n
\]

where the given time series \((X_i, Y_i)\) are generated from this model, which is with no explicit error term. The unobserved coefficients \((A_i, B_i)\) are assumed to be independent random vectors generated from some unknown distribution with bivariate density \(f_{AB}(a, b)\). The task is to estimate \(f_{AB}\) from the observed \((X_i, Y_i)\)'s, and the problem of estimating the regression function is often called deconvolution in the engineering literature. Feuerverger and Vardi showed that the nonparametric likelihood function for the RCR problem is the same as that of a positron emission
tomography (PET) image reconstruction problem with a suitably specified conditional probability of detection function. With this relation, many of the statistical methods for image reconstruction in PET, like convolution backprojection, Bayes methods, penalized likelihood, and smoothed EM algorithms can be modified for the RCR problem.

RCR is not only of interest in econometrics, but has been applied in other fields as well. Corbett and et al. (2003) built a RCR model to estimate growth curve parameters for the serum glucose trait. The resulting parameters were then used as phenotypes for a variance-components based linkage analysis method. It is shown that the simulated time series is nearly ideal for examination using the RCR model, and the RCR model can virtually recover the model used to simulate the glucose level time series. The RCR growth parameter is proved much more powerful at discovering the location of loci affecting the growth curve for serum glucose levels than the two time-point slope phenotype.

2.3.7 Discrete Fourier Transform

Discrete Fourier transform (DFT) is very useful because they can reveal periodicities in input data as well as the relative strengths of any periodic components (Weisstein 2009). DFT is a specific kind of Fourier transform, with discrete inputs which are often created by sampling a continuous function, like the human voice. The model decomposes a sequence of values into components of different frequencies. Mathematically, the DFT is defined by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi ink}{N}} \quad k = 0, \ldots, N-1$$

where the time series $x_0, \ldots, x_{N-1}$ are complex numbers. As there are $N$ outputs $X_k$ and each output requires a total of $N$ terms, it takes $O(N^2)$ operations as a whole, where $O$ denotes the upper bound of the computational complexity. The inverse discrete Fourier transform (IDFT) is defined by

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi ink}{N}} \quad n = 0, \ldots, N-1$$

As the inputs for the DFT is a finite sequence of numbers, the model is very suitable for processing information stored in computers, and is widely used in signal processing to study the frequencies in the sample signal. In practice, the DFT is commonly computed much more efficiently using a fast Fourier transform (FFT) algorithm. The FFT algorithms only require $O(N\log N)$ operations. Currently the most common FFT is the Cooley–Tukey algorithm, which is a divide and conquer algorithm that recursively breaks down a DFT of any composite size $N = N_1 N_2$ into many smaller DFTs of sizes $N_1$ and $N_2$ (Cooley and Tukey 1965).
Applied Time Series Analysis and Innovative Computing
Ao, S.-I.
2010, XIV, 112 p., Hardcover