Singular Spectrum Analysis

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

No Kang Myung

2009
© Copyright by
No Kang Myung
2009
The thesis of No Kang Myung is approved.

________________________________________________________________________
Rick Paik Schoenberg

________________________________________________________________________
Christou Nicolas

________________________________________________________________________
Jan de Leeuw, Committee Chair

University of California, Los Angeles
2009
To my parents . . .
# Table of Contents

1 Introduction .................................................................................. 1

2 Data ............................................................................................. 3

3 Basic Singular Spectrum Analysis algorithm ............................... 5
   3.1 Decomposition ........................................................................... 5
      3.1.1 Embedding .......................................................................... 5
      3.1.2 Singular Value Decomposition ............................................ 6
   3.2 Reconstruction .......................................................................... 6
      3.2.1 Grouping ............................................................................... 6
      3.2.2 Diagonal averaging ............................................................. 7

4 SSA Conditions for Graphical Inference ................................... 9
   4.1 Selection of SSA Parameters: Window length effects ......... 9
   4.2 Selection of SSA Parameters: Grouping effects ................. 10
   4.3 Extracting Condition: Weak separability ............................ 10
   4.4 Single centering ...................................................................... 11

5 Graphical Inference ....................................................................... 12
   5.1 $CO$, $SO_2$, and $NO_2$ Mean ................................................ 12
   5.2 Ozone Mean and Temperature ............................................... 13
   5.3 Cardiovascular Death (CVD) ................................................. 14
   5.4 The conclusion of graphical inference ................................. 15

6 Multi-channel Singular Spectrum Analysis ............................... 27
   6.1 The MSSA algorithm ............................................................. 27
6.1.1 Embedding ............................................. 27
6.1.2 SVD ................................................. 29
6.1.3 Grouping and diagonal averaging ................. 30

7 Conclusions ...................................................... 31

References ......................................................... 32
**List of Figures**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>The CO trend reconstructed by 1-2 eigentriples when $L = 365$</td>
<td>16</td>
</tr>
<tr>
<td>5.2</td>
<td>The CO trend with 1-2 eigentriples for 3 years</td>
<td>17</td>
</tr>
<tr>
<td>5.3</td>
<td>half yearly (16-17 eigentriples): monthly (4, 7, 21, and 24 eigentriples)</td>
<td>17</td>
</tr>
<tr>
<td>5.4</td>
<td>The SO2 trend reconstructed by 1-2 eigentriples when $L = 365$</td>
<td>18</td>
</tr>
<tr>
<td>5.5</td>
<td>The SO2 trend reconstructed by 1-2 eigentriples for 3 years: monthly (24 eigentriples)</td>
<td>18</td>
</tr>
<tr>
<td>5.6</td>
<td>The NO2 trend reconstructed by 1-2 eigentriples when $L = 365$</td>
<td>19</td>
</tr>
<tr>
<td>5.7</td>
<td>The NO2 trend reconstructed by 1-2 eigentriples for 3 years</td>
<td>20</td>
</tr>
<tr>
<td>5.8</td>
<td>half-yearly (24 eigentriple): monthly (11 eigentriple)</td>
<td>20</td>
</tr>
<tr>
<td>5.9</td>
<td>The Ozone trend reconstructed by 1-2 eigentriples when $L = 365$</td>
<td>21</td>
</tr>
<tr>
<td>5.10</td>
<td>The trend with 1-2 eigentriples for 3 years</td>
<td>22</td>
</tr>
<tr>
<td>5.11</td>
<td>half-yearly (3-4 eigentriples): quarterly (9-10 eigentriples): monthly (5-6 eigentriples)</td>
<td>22</td>
</tr>
<tr>
<td>5.12</td>
<td>The temperature trend reconstructed by 1-2 eigentriples when $L = 365$</td>
<td>23</td>
</tr>
<tr>
<td>5.13</td>
<td>The trend with 1-3 eigentriples for 3 years</td>
<td>24</td>
</tr>
<tr>
<td>5.14</td>
<td>half-yearly(4-5 eigentriples): quarterly (6 eigentriple): monthly (12-13 eigentriples)</td>
<td>24</td>
</tr>
<tr>
<td>5.15</td>
<td>The CVD trend reconstructed by 1st eigentriple when $L = 365$</td>
<td>25</td>
</tr>
<tr>
<td>5.16</td>
<td>The CVD smoothing reconstructed by 1st and 2nd eigentriples when $L = 365$</td>
<td>25</td>
</tr>
<tr>
<td>5.17</td>
<td>The trend with 1-3 eigentriples for 3 years</td>
<td>26</td>
</tr>
<tr>
<td>5.18</td>
<td>half-yearly(4-5 eigentriples): quarterly (6-7 eigentriples)</td>
<td>26</td>
</tr>
</tbody>
</table>
LIST OF TABLES

4.1 The effect of window length .......................... 10
Abstract of the Thesis

Singular Spectrum Analysis

by

No Kang Myung

Master of Science in Statistics

University of California, Los Angeles, 2009

Professor Jan de Leeuw, Chair

Air pollution data recorded in the U.S. are multivariate time series of many cities. Statistics analysis was used and tried to analyze multivariate air pollution data and discovered that the time series methods fit to the air pollution data. The Singular Spectrum Analysis (SSA) method was developed as the new time series method since 1970s and they are still growing mathematically. SSA regards as a model-free approach because SSA decomposes an original time series to trend, seasonal, semi-seasonal, and white noises according to the singular value decomposition (SVD) \[4\]. The new decomposed series help us to understand the trend of the original time series and to extract seasonal or monthly components and white noises. In recent years, SSA did not only apply in the analysis of climatic and geophysical time series, but also in the analysis of social science and economic time series. With these diverse areas, air pollution time series in Los Angeles is applied to SSA. As a methodology, the theories of SSA will be discussed in the paper.
CHAPTER 1

Introduction

Singular Spectrum Analysis (SSA) is a recent and alternative time series method. Broomhead and King (1986a, 1986b) contributed to the birth of SSA because they show that the singular value decomposition (SVD) is effective in reducing noises. Simultaneously, some research groups in the UK, USA and Russia developed SSA main ideas. Vautard et al. (1992), Ghil and Taricco (1997), Allen and Smith (1986), Danilov (1997), Yiou et al. (2000) published several papers dealing with methodologies and applications of SSA [2]. Elsner and Tsonis (1996) published a time series book whose title is *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. This book became the primary introduction to SSA.

In Elsner and Tsonis’ book, SSA is made of two terms, Singular and Spectrum. The term singular comes from the spectral decomposition or eigen-decomposition of a matrix T; eigenvalues are used for decomposing each column of T. The term spectrum is defined to add a set of eigenvalues after spectral decomposition. More interestingly, SSA is analogous to multivariate data analysis so that it can be useful to reduce the dimensionality of T [7].

The data that apply in SSA are from the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. The Biostatistics department maintains the Internet-based Health & Air Pollution Surveillance System (IHAPPS) [3]. The IHAPPS collected data from the U.S. Environmental Protection Agency, the National Center for Health Statistics, the National Climatic Data Center, and the U.S. Census Bureau. IHAPPS provides us the National Mortality Morbidity Air Pollution Study (NMMAPS) data [3].

The basic concept of SSA consists of four steps: embedding, singular value de-
composition (SVD), grouping, and diagonal averaging. The basic SSA algorithm partitioned an initial time series into new time series which consist of the trend, seasonal, monthly components, and white noises. In chapter three, we will explain a detailed methodology of these four steps.

With the basic SSA algorithm, we will analyze air pollutants, temperature, and cardiovascular death in the NMMAPS data which are complex time series. In addition we will discuss convenient and limited aspects of an extended SSA technique which is called Multi-Channel Singular Spectrum Analysis (MSSA).
CHAPTER 2

Data

IHAPPS stands for the Internet-based Health & Air Pollution Surveillance System operated by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health [3]. The purpose of IHAPPS is to monitor air pollution data and study the associative effects that air pollution has on mortality and morbidity in the United States [3]. IHAPPS has provided the National Mortality Morbidity Air Pollution Study (NMMAPS) data.

The NMMAPS data have included air pollution, weather, and death causes of 108 cities in the U.S. between January 1, 1987 and December 31, 2000. Air pollution data, daily mortality counts, weather data, and census data come from the AirData database collected by the U.S. Environmental Protection Agency, the National Center for Health Statistics, the National Climatic Data Center, and the United States Census Bureau.

The NMMAPS data dimensions in the city of Los Angeles have 15342 rows, which are daily data for fourteen years, and 131 columns, which are variables. There are three age categories: under 65 years old, 65 to 75 years old, and over 75 years old. Thus each group has 5114 daily time series data with the same air pollution and weather data, but with different mortality and morbidity data according to each age group.

In the NMMAPS data, air pollution data have six main primary pollutants; Particular Matter 10 ($PM_{10}$), Particular Matter 2.5 ($PM_{2.5}$), Ozone ($O_3$), Sulfur Dioxide ($SO_2$), Carbon Monoxide ($CO$), and Nitrogen Dioxide ($NO_2$). Temperature, dew point temperature, and relative humidity belong to weather data. There are main causes of people deaths; total non-accidental deaths, cardiovascular deaths (CVD), respiratory deaths, chronic obstructive pulmonary disease (COPD), accidental deaths [3].

The sources of air-pollutants have come from mostly human activities: driving
a vehicle and an air-plain, and running a factory [10]. $O_3$ is called bad and lower level ozone which is a pollutant, and a part of smog. The incomplete burning of fuels such as gas, oil, coal, and wood produces $CO$, $NO_2$, and $SO_2$. $CO$ and $NO_2$ are also deadly tasteless, odorless and colorless gases [10]. Exposure to these air pollutants causes people to have acute and chronic bronchitis, asthma, respiratory infections, and heart diseases. Furthermore, air pollutants lead people to respiratory and cardiovascular death.
CHAPTER 3

Basic Singular Spectrum Analysis algorithm

The basic SSA algorithm has two stages: decomposition and reconstruction. The decomposition stage requires embedding and singular value decomposition (SVD). Embedding decomposes the original time series into the trajectory matrix; SVD turns the trajectory matrix into the decomposed trajectory matrices which will turn into the trend, seasonal, monthly components, and white noises according to their singular values. The reconstruction stage demands the grouping to make subgroups of the decomposed trajectory matrices and diagonal averaging to reconstruct the new time series from the subgroups.

3.1 Decomposition

3.1.1 Embedding

The first step in the basic SSA algorithm is the embedding step where the initial time series change into the trajectory matrix. Assume we have a daily time series of length \( N \) without any missing values, and let the time series be \( X = \{x_1, \ldots, x_N\} \). In the embedding step, we choose window length \( L \), where \( 2 < L < \frac{N}{2} \) to embed the initial time series \([1]\). We map the time series \( X \) into the \( L \) lagged vectors, \( X_i = \{x_i, \ldots, x_{i+L-1}\} \) for \( i = 1, \ldots, K \), where \( K = N - L + 1 \) \([4]\). We make the trajectory matrix \( T_X \) with \( X_i \) which is each row of \( T_X \) for \( i = 1, \ldots, K \). The trajectory matrix \( T_X \) with \( L \times K \) dimensions has one main property; the trajectory matrix \( T_X \) shows that cross-diagonals of \( T_X \) is \( x_{j+i-1} = x_{i+j-1} \) \([4]\). Thus \( T_X \) is written as
\[
T_{i,j} = \begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
X_K
\end{pmatrix} = \begin{pmatrix}
x_1 & x_2 & \ldots & x_L \\
x_2 & x_3 & \ldots & x_{L+1} \\
\vdots & \vdots & \ddots & \vdots \\
x_K & x_{K+1} & \ldots & x_N
\end{pmatrix}
\]

### 3.1.2 Singular Value Decomposition

After the embedding step, we apply SVD to the trajectory matrix \(T_X\) and obtain the decomposed trajectory matrices \(T_i\) for \(i = 1, \ldots, L\). The outcome of applying SVD to \(T_X\) decomposes \(T_X = UDV^T\) called by eigentriples [1]. \(U_i\) for \(1 < i < L\) is a \(K \times L\) orthonormal matrix. \(D_i\) for \(1 < i < L\) is a diagonal matrix order of \(L\). \(V_i\) for \(1 < i < L\) is a \(L \times L\) square orthonormal matrix [4]. In this step, \(T_X\) has \(L\) many singular values which are \(\sqrt{\lambda_1} > \sqrt{\lambda_2} > \ldots, > \sqrt{\lambda_L}\). Thus the \(i\)th eigentriple of \(T_i\) can be written as \(U_i \times \sqrt{\lambda_i} \times V_i^T\) for \(i = 1, 2, \ldots, d\), in which \(d = \max (i : \sqrt{\lambda_i} > 0)\). So the trajectory matrix \(T_X\) can be denoted as

\[
T_X = T_1 + T_2 + \ldots + T_d \\
= U_1\sqrt{\lambda_1}V_1^T + \ldots + U_d\sqrt{\lambda_d}V_d^T \\
= \sum_{i=1}^{d} U_i\sqrt{\lambda_i}V_i^T
\]

We can calculate the ratio of each eigenvalue \(\lambda_i / \sum_{i=1}^{d} \lambda_i\) since \(\|T_X\|^2 = \sum_{i=1}^{d} \lambda_i\) and \(\|T_i\|^2 = \lambda_i\) for \(i = 1, \ldots, d\) [9]. The ratio of each eigenvalue \(\lambda_i / \sum_{i=1}^{d} \lambda_i\) is the contribution of the matrix \(T_i\) to \(T_X\).

### 3.2 Reconstruction

#### 3.2.1 Grouping

The grouping step of the reconstruction stage is decompose the \(L \times K\) matrix \(T_i\) into subgroups according to the trend, seasonal, monthly components, and white noises. The
The grouping step of the reconstruction stage is a partition of the set of indices \( \{1, \ldots, d\} \) into the collection of \( m \) disjoined subsets of \( I = \{I_1, \ldots, I_m\} \) \cite{[1]}. Thus \( T_I \) corresponds to the group \( I = \{I_1, \ldots, I_m\} \) \cite{[1]}. \( T_I \) is a sum of \( T_j \), where \( j \in I_i \). So \( T_X \) can be expanded as

\[
T_X = \underbrace{T_1 + \ldots + T_L}_{\text{Grouping}}
\]

\[
= T_{I_1} + \ldots + T_{I_m}.
\]

To understand the grouping step, for example we assume that there are only two groups of eigentriplets of the trajectory matrix \( T_X \): two groups are \( T_L \) and \( T_R \). Let the entire set \( I = \{1, \ldots, d\}, R \cup L = I \), but \( R \) is not a subset of \( L \) \cite{[1]}. \( T_I \) is

\[
T_I = \sum_{i \in I} \sqrt{\lambda_i} U_i V_i^T
\]

and we can calculate \( T_L = T_I - T_R \) under the assumption of weak separability. Thus \( T_L \) can be written as

\[
T_L = \sum_{i \in L} \sqrt{\lambda_i} U_i V_i^T
\]

Weak separability is necessary to support grouping and diagonal averaging. At this moment, we skip its mathematical explanation and discuss it in chapter four.

### 3.2.2 Diagonal averaging

In the basic SSA algorithm, the diagonal averaging step is to transform the grouped matrices \( T_{I_i} \) into a new time series of length \( N \) \cite{[1]}. We obtain the time series \( \tilde{T}^{(i)} \) from averaging of the corresponding diagonals of the matrix \( T_{I_i} \) \cite{[1]}

\[
\tilde{T}_{i,j} = \begin{cases} 
\frac{1}{s-1} \sum_{j=1}^{s-1} x_{j,s-j} & \text{for } 2 \leq s \leq L \\
\frac{1}{L} \sum_{j=1}^{L} x_{j,s-j} & \text{for } L + 1 \leq s \leq K + 1 \\
\frac{1}{N-s+2} \sum_{j=s-K}^{N-s+2} x_{j,s-j} & \text{for } K + 2 \leq s \leq N + 1 
\end{cases}
\]
Let the Hankelization operator $H$ be averaging of the corresponding diagonals of the matrix $T_{II}$ for $i = 1, \ldots, m$. Hankelization procedure uses the Hankelization operator $H$ to transform $T_{II}$ into $\tilde{X}^{(i)} = HT_{II}$ for $i = 1, \ldots, m$ [1]. Under the assumption of weak separability the initial time series $X$ can be reconstructed by

$$X = \tilde{X}^{(1)} + \tilde{X}^{(2)} + \ldots + \tilde{X}^{(M)}$$

We can assert $\tilde{X}^{(1)}$ always becomes the trend; however, seasonal and monthly components do not follow the order of $\sqrt{\lambda_1} > \sqrt{\lambda_2}, \ldots, > \sqrt{\lambda_M}$. 
CHAPTER 4

SSA Conditions for Graphical Inference

4.1 Selection of SSA Parameters: Window length effects

In the embedding step, window length $L$ is a main parameter to determine the dimension of the trajectory matrix $T_X$. Let $X$ be the initial time series $\{x_1, x_2, \ldots, x_N\}$ with no missing values of length $N$. The selection of window length $L$ decides to map the time series into $L$ lagged vector $X_i = \{x_i, \ldots, x_{i+L-1}\}$ for $i = 1, \ldots, K$, where $K = N - L + 1$ [4]. The range of window length $L$ is $2 < L < \frac{N}{2}$. Because window length $L$ corresponds to $K = N - L + 1$ in the trajectory matrix $T_X$, $L$ is no longer necessary to be larger than $\frac{N}{2}$ [1].

Suppose that we have a complex time series in the NMMAPS data for three years. As window length $L$ increases up to $\frac{N}{2}$, the decomposition of the time series becomes more detailed. Since the time series is complex, the maximum rank of the trajectory matrix $T_X$ becomes $\frac{N}{2}$. However, we have to be careful to choose the window length $L$ with the periodic time series. When the data related to air pollutions are periodic, selection of window length $L$ can be neither the maximum nor the minimum of window length $L$. If window length $L$ is close to $\frac{N}{2}$, it can make the trajectory matrix $T_X$ to overlap its rows and columns. If window length $L$ is relatively small, it can cause the trajectory matrix $T_X$ to produce the improper decomposition of the data because the trend contains seasonal or monthly components [1].

The table 4.1 is a summary table for the effects of window length in SSA. The larger we increase window length, the smaller the first singular eigentriple includes the singular value $\sqrt{\lambda_1}$. Therefore, if we select proper window length $L$, the trend becomes a smooth or regular pattern. If we decrease window length $L$, the trend reverses, and
becomes rough or shows an irregular pattern because it contains other components.

<table>
<thead>
<tr>
<th>Window Length</th>
<th>Rank of $T_X$</th>
<th>$\sqrt{\lambda_1}$</th>
<th>Main Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>smooth</td>
</tr>
<tr>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>rough</td>
</tr>
</tbody>
</table>

Table 4.1: The effect of window length

4.2 Selection of SSA Parameters: Grouping effects

In the reconstruction stage, grouping as the main parameter plays a significant role to determine the trend, smooth, half-yearly, quarterly, monthly series and white noises for the graphical inference. After every decomposition step is done with certain window length, we group the SVD terms of the trajectory matrix $T_X$ to present properly separated series of the initial series [1].

It is important to know how we find proper sub-groups in the grouping step. The given conditions are $L$ many eigentriples from SVD and $\tilde{X}^{(i)}$ for $i = 1, \ldots, d$ from diagonal averaging. The SVD terms provide each eigentriple with the ratio of a singular value, and we can depict the plot of Hankelization procedure of each $\tilde{X}^{(i)}$. Frankly speaking, looking at several plots of leading eigentriples helps to decide the trend or smoothing, and other components.

4.3 Extracting Condition: Weak separability

The Singular Value Decomposition (SVD) step divides the trajectory matrix into $L$ many decomposed trajectory matrices. We can make at most $L$ many matrices after the grouping step if $d = L$, where $d = \max(i : \sqrt{\lambda_i} > 0)$. In the SVD and diagonal averaging step, we need the assumption that extracting the trend and other components from the original time series is weakly separable [1]. We call the assumption weak separability. Suppose that we have only two time series: $F^{(1)}$ and $F^{(2)}$. Covariance of $(F^{(1)}, F^{(2)})$ can be written by
\[ \rho_{12} = \frac{(F^{(1)}, F^{(2)})}{\|F^{(1)}\| \|F^{(2)}\|} \]

where \( \|F^{(i)}\| = \sqrt{(F^{(i)}, F^{(i)})}, i = 1, 2 \) [1].

The series \( F^{(1)} \) and \( F^{(2)} \) become \( w \)-orthogonal if \( (F^{(1)}, F^{(2)})_w = 0 \) [1]. If the time series \( F \) is \( w \)-orthogonal, then we can calculate \( F = F^{(1)} + F^{(2)} \). Under \textit{weak separability}, we apply SSA to air pollution data and extract the trend or components from the initial data. Then all extractions of the trend or components from initial data is based on \textit{weak separability} in chapter five.

### 4.4 Single centering

Suppose that we have a time series \( X_N \) with length \( N \). We make generally the trajectory matrix \( T_X \) with window length \( L \) and \( K = N - L + 1 \). Single centering is the step where we subtract the mean of the column of \( T_X \) from each element of a column of \( T_X \). Single centering is denoted as

\[
X = A_1 + \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T = A_1(X) + \sum_{i=1}^{d} X_i
\]

In the applications of SSA in chapter five, the SSA-MTM Toolkit subtracts the mean from initial data like single centering. Therefore, the SSA-MTM Toolkit shows a new time series which is oscillating around zero. \( O_3 \), \( SO_2 \), \( CO \), and \( NO_2 \) are similar to single centered data provided from the NMMAPS data, but CVD and temperature are not single centered data. Therefore, there are differences in plotting CVD and temperature data between the program R and the SSA-MTM Toolkit.
CHAPTER 5

Graphical Inference

Applying SSA to CVD, Temperature, $O_3$, $SO_2$, $CO$, and $NO_2$ data, we definitely extract the trend and smoothing, and we may extract half-yearly, quarterly, or monthly harmonic series. After we decompose the initial series into these separate series corresponding to their eigenvalues, the decomposed plots help us to understand what each time series looks like. With the illustrated conditions in the previous chapter, we plot the trend or smoothing on the initial time series and select proper half-yearly, quarterly, or monthly harmonic series. The time period of the trend or smoothing is equal to the entire time period of the data. The time periods of half-yearly, quarterly, and monthly series are six, four, and one months.

In the book of N. Golyandina et al. (2001), the definition of a trend is described as a slowly varying component of a time series which does not contain oscillatory components [1]. Smoothing extracts all eigentriples pertaining to a slowly varying and non-oscillatory part of the series and all eigentriples corresponding to slowly varying and constantly oscillatory parts of the series [1]. In addition, we will use a smooth trend because the trend contains its own trend plus the oscillating pattern.

5.1 CO, SO$_2$, and NO$_2$ Mean

We put $CO$, $SO_2$, and $NO_2$ mean data at the same section in order to interpret them graphically. $CO$, $SO_2$, and $NO_2$ mean values and their smooth trends are depicted in Figure 5.1, Figure 5.4, and Figure 5.6. Figure 5.1 shows that the amplitude of $CO$ mean values was reduced for fourteen years and the trend is reconstructed with 1-2 eigentriples. We cut off the first three years of series and reapply them to SSA using R. So
Figure 5.2 is the plot where three years’ worth of daily data of CO mean values and its trend are described. The trend is a kind of an average curve of CO mean values; it shows a pattern that the trend is lower in the winter and higher in the summer. We find half-yearly components (16-17 eigentriples), and monthly components (4, 7, 21, and 24 eigentriples) in figure 5.3, whereas any quarterly component is not found. Half-yearly components show a constant variation during three years, and monthly components are much wider than the summer variation of monthly components.

The trend of SO$_2$ mean in figure 5.4 lose its oscillating pattern because the yearly pattern of SO$_2$ mean had been weak approximately after 1996. We apply the first three year (1987-1989) series which have a strong yearly pattern to SSA. The left plot of figure 5.5 shows SO$_2$ mean and its trend; the trend is reconstructed by 1-2 eigentriples. The right plot of figure 5.5 is monthly series of SO$_2$ mean reconstructed by only the 24th eigentriple because we cannot find any proper half-yearly and quarterly components.

Figure 5.6 describes NO$_2$ mean data and its trend; the amplitude of NO$_2$ mean has not changed much and the smooth trend (1-2 eigentriples) shows the constant oscillating pattern for fourteen years. Using R, NO$_2$ data for three years and its trend (1-2 eigentriples) are depicted in figure 5.7. NO$_2$ mean data also have higher variations in winter than in summer. A half-yearly component (24 eigentriple) on the left plot and a monthly component (11 eigentriple) on the right plot are shown in figure 5.8.

### 5.2 Ozone Mean and Temperature

In figure 5.9 and 5.12, temperature and O$_3$ show similar patterns in which they are higher in summer than in winter. On the other hand, CO, SO$_2$, and NO$_2$ mean data are higher in winter than in summer. The O$_3$ mean data and the amplitude of the trend (1-2 eigentriples) gradually reduced for fourteen years in figure 5.9. The three years O$_3$ mean data and its trend (1-2 eigentriples) are depicted in figure 5.10. There are three plots in figure 5.11: half-yearly (3-4 eigentriples) harmonic components on the left plot, quarterly (9-10 eigentriples) harmonic components on the middle plot, and monthly (5-6 eigentriples) harmonic components on the right plot in the three year time scale. The
three harmonic components of $O_3$ indicate their periodic patterns well. According to the eigenvalues of components, half-yearly components have the highest impact on $O_3$; monthly components have the second highest impact on $O_3$, and quarterly components have the least impact on $O_3$.

Temperature data have only eleven years’ worth of daily data and temperature data from 1997 to 2000 are missing data. Temperature data are not around zero like $O_3$, $SO_2$, $CO$, and $NO_2$. Figure 5.12 ran in the Toolkit shows temperature and the trend (1-2 eigentriples) around zero for eleven years. Using R displays the initial times series and the trend (1-3 eigentriples) for three years in Figure 5.13. Figure 5.14 consists of three plots in which half-yearly (4-5 eigentriples) harmonic components on the left plot, quarterly (6 eigentriple) harmonic component on the middle plot, and monthly (12-13 eigentriples) harmonic components on the right plot.

5.3 Cardiovascular Death (CVD)

An entire pattern of CVD data is slightly higher in winter than in summer. The result of using the Toolkit displays CVD mean data and the trend (1 eigentriple) in figure 5.15. Among six variables, CVD shows the trend defined in N. Golyandina’s book at page 55. CVD mean data and the smooth trend (1-2 eigentriple) are depicted in figure 5.16 where CVD has a seasonal oscillating pattern and CVD was decreasing for fourteen years through the smooth trend.

The initial CVD data and the smooth trend (1-3 eigentriple) are depicted in figure 5.17 where a weak seasonal pattern is shown unlike the trend in figure 5.16. Figure 5.18 shows half-yearly (4-5 eigentriples) harmonic components on the left plot and quarterly (6-7 eigentriples) harmonic components on the right plot. Half-yearly components affect CVD more than quarterly components because half-yearly eigentriples come before monthly eigentriples.
5.4 The conclusion of graphical inference

The CVD trend depicted in figure 5.15 follows the trend definition well. Except figure 5.15, the other trends in chapter five show the smooth trend because they include smooth oscillatory components. Watching the trend plots, we figure out that distinguishing the trend and smoothing may not be appropriate in the complex and complicated time series data.

We cannot make a certain model to extract proper harmonic components in SSA because SSA just uses the model-free method based on the SVD. Applying SSA to \( SO_2 \) mean data shows a poor example of the result of the model-free model in figure 5.5 where we cannot find any half-yearly and quarterly harmonic components. However, we can extract proper harmonic series from \( O_3 \) mean and temperature using SSA.
Figure 5.1: The CO trend reconstructed by 1-2 eigentriples when $L = 365$
Initial series of CO and its yearly trend

Figure 5.2: The CO trend with 1-2 eigentriples for 3 years

Figure 5.3: half yearly (16-17 eigentriples): monthly (4, 7, 21, and 24 eigentriples)
Figure 5.4: The SO2 trend reconstructed by 1-2 eigentriples when L = 365

Figure 5.5: The SO2 trend reconstructed by 1-2 eigentriples for 3 years: monthly (24 eigentriples)
Figure 5.6: The NO2 trend reconstructed by 1-2 eigentriples when $L = 365$
Initial series of NO2 and its yearly trend

Figure 5.7: The NO2 trend reconstructed by 1-2 eigentriples for 3 years

Figure 5.8: half-yearly (24 eigentriple): monthly (11 eigentriple)
Figure 5.9: The Ozone trend reconstructed by 1-2 eigentriples when $L = 365$
Figure 5.10: The trend with 1-2 eigentriples for 3 years

Figure 5.11: half-yearly (3-4 eigentriples): quarterly (9-10 eigentriples): monthly (5-6 eigentriples)
Figure 5.12: The temperature trend reconstructed by 1-2 eigentriples when $L = 365$
Figure 5.13: The trend with 1-3 eigentriples for 3 years

Figure 5.14: half-yearly(4-5 eigentriples): quarterly (6 eigentriple): monthly (12-13 eigentriples)
Figure 5.15: The CVD trend reconstructed by 1st eigentriple when L = 365

Figure 5.16: The CVD smoothing reconstructed by 1st and 2nd eigentriples when L = 365
Initial series of CVD and its yearly trend

Figure 5.17: The trend with 1-3 eigentriples for 3 years

Half-yearly series of CVD
Quarterly series of CVD

Figure 5.18: half-yearly(4-5 eigentriples): quarterly (6-7 eigentriples)
CHAPTER 6

Multi-channel Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) can be extended to Multi-channel Singular Spectrum Analysis (MSSA) for a multivariate time series of vectors at concurrent moments at a same location or at different locations. MSSA was theoretically proposed by Broomhead and King [1986b] to explain nonlinear dynamics. MSSA is used to approach to systems of ordinary or partial differential equations [5]. The major areas of MSSA are multivariate statistics, multivariate geometry, and dynamical systems and signal procession.

6.1 The MSSA algorithm

6.1.1 Embedding

In order to explain MSSA, we need to include vector or location variables $m = 1, \ldots, M$, except window length $L$ between 2 and $T/2$. $M$ can be either multi-variables for one location or multi-channels for various locations. MSSA uses two parameters in describing each data as $\{X_{l,m} : l = 1, \ldots, L; m = 1, \ldots, M\}$ [5]. We assume that $M$ is equally spaced or timely intervals and is stationary in the weak sense [5].
Figure 6.1: MSSA [5]

Figure 6.1 illustrates schematically the distribution of the multi-channel data for MSSA; axes of t, x, and s correspond to time N, window length L, and multi-channel M [5]. There are special cases in MSSA. If M = 1, MSSA becomes SSA because there is no other vectors or channels. If L = 1, MSSA becomes PCA like multivariate statistics.

When we decompose the grand covariance matrix $C_X$, we previously require two more steps: the each-channel’s trajectory matrix $X_m$ and the full augmented trajectory matrix called the multi-channel trajectory matrix also $\tilde{X}$. We form the each-channel’s trajectory matrix $\tilde{X}_m$ with $\{X_{l,n} : l = 1, \ldots, L; m = 1, \ldots, M\}$

$$
\tilde{X}_m = \begin{pmatrix}
  x_{m,1} & x_{m,2} & \cdots & x_{m,L} \\
  x_{m,2} & x_{m,3} & \cdots & x_{m,L+1} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{m,K} & x_{m,K+1} & \cdots & x_{m,N}
\end{pmatrix}, \ 1 \leq m \leq M
$$

and then the full augmented trajectory matrix $\tilde{X}$ [6]. The full augmented trajectory matrix $\tilde{X}$ is defined as

$$
\tilde{X} = \begin{pmatrix}
  \tilde{X}_1 \\
  \tilde{X}_2 \\
  \vdots \\
  \tilde{X}_M
\end{pmatrix}, \ 1 \leq m \leq M
$$
The last step of decomposition in MSSA calculates the grand covariance matrix $C_X$ as

$$C_X = \frac{1}{N'} \tilde{X}^T \tilde{X} = \begin{pmatrix} C_{1,1} & C_{1,2} & \ldots & C_{1,M} \\ C_{2,1} & C_{2,2} & \ldots & C_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ C_{M,1} & C_{M,2} & \ldots & C_{M,M} \end{pmatrix}, 1 \leq m \leq M$$

The dimension of the grand covariance matrix $C_X$ becomes $(L \times K) \times (L \times K)$. Each block of $C_X$ is described as

$$C_{m,m'} = \frac{1}{N'} \tilde{X}_{m}^T \tilde{X}_{m'}$$

Each entry of $C_X$ is written as

$$(C_{m,m'})_{j,j'} = \frac{1}{N'} \sum_{t=1}^{N'} X_{m}(t + j - 1)X_{m'}(t + j' - 1)$$

where $N'$ is a normalized factor depending on the range [10]. $N'$ is

$$N' = \min(N, N + j - j') - \max(1, 1 + j - j') + 1$$

$C_X$ is symmetric because $C_{m,m'} = (C_{m',m})^T [10]$.

### 6.1.2 SVD

After embedding multi-channel time series, we decompose the grand block matrix by applying the principle component analysis (PCA) instead of the SVD. We summarize the PCA properties using proposition 4.1 in the book of N. Golyandina. et al. Suppose there are vector $U_i$ and $V_i$.

1. Let $1 \leq i, j \leq d$. Then $(V_i, V_j) = 0$ for $i \neq j$ and $\|V_i\|$

2. $V_i$ is an eigenvector of the matrix $X^TX$ corresponding to an eigenvalue $\lambda_i$.

Then $X$ can be written as:

$$X = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T$$

Since $C_X$ is followed by proposition 4.2
\[ C_X = X X^T = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T \]

\[ \sum_{j=1}^d \sqrt{\lambda_j} U_j^T V_j = \sum_{i=1}^d \lambda_i V_i V_i^T \]

The diagonals where \( C_{i,j} \) for \( i = j \) of the grand covariance matrix are decomposed by using the SVD. If we ignore off-diagonals of the grand covariance matrix, MSSA just combines several time series and apply SSA to them. When we apply the PCA or the SVD to off-diagonals where \( C_{i,j} \) for \( i \neq j \) of the grand covariance matrix, we cannot use the PCA and the SVD to decompose. Thus we need a new method to decompose off-diagonals because \( C_{i,j} \) for \( i \neq j \) becomes a totally different matrix. Furthermore, if \( m \) is not equal to \( m' \), and they are independent, \( C_{m,m'} \) becomes useless or meaningless. This part is one of the weak points of MSSA not to interpret off-diagonals, which means we need to develop an alternative method to calculate the off-diagonals.

6.1.3 Grouping and diagonal averaging

The steps of grouping and diagonal averaging follow mostly the steps of SSA because of the limitation of MSSA discussed in the previous subsection. One block where \( C_{i,j} \) for \( i = j \) of the grand covariance matrix follow the steps of SSA grouping and diagonal averaging. However, other blocks where \( C_{i,j} \) for \( i \neq j \) are ignored in two steps of MSSA because they cannot be applied to the PCA or the SVD.

MSSA has an advantage in depicting a variable located at \( M \) many different locations at one plot. Suppose we have \( SO_2 \) and \( CO \) mean data depicted like figure 5.1 and 5.4 at \( M \) multi-channels. We should plot the trends of \( SO_2 \) and \( CO \) mean data at separate plots because the ranges of \( SO_2 \) and \( CO \) mean data are extremely different.
CHAPTER 7

Conclusions

1. SSA, a new time series technique, suggests graphical analysis in the study of time series. SSA is a good method to extract the trend and other components for simple data structures. SSA applies to complex data structures, and then we may extract proper components except the trend.

2. SSA is based on model-free techniques so that we cannot be active to extract the eigentriples, but can be passive to select the proper eigentriples corresponding to the eigentriple plots.

3. Program R is not the proper software to run large data. The SSA-MTM Toolkit well made software run in Mac and Linux is much faster than Program R because UCLA SSA-MTM group has developed and updated it since 2000.

4. The NMMAPS data are complex time series with strong periodic yearly patterns so that their trends become oscillating smooth trends except CVD which shows a weak periodic yearly pattern.

5. We found out the limitation to decompose off-diagonals of the grand covariance matrix in MSSA. We need an effort to solve the limitation.
REFERENCES


