INFERENCE AND VISUALIZATION OF PERIODIC SEQUENCES

A Dissertation

by

YING SUN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2011

Major Subject: Statistics

INFERENCE AND VISUALIZATION OF PERIODIC SEQUENCES

A Dissertation

by

YING SUN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Jeffrey D. Hart
	Marc G. Genton
Committee Members,	Raymond J. Carroll
	Kenneth P. Bowman
Head of Department,	Simon J. Sheather

August 2011

Major Subject: Statistics

ABSTRACT

Inference and Visualization of Periodic Sequences. (August 2011)

Ying Sun, M.S., Tsinghua University

Co–Chairs of Advisory Committee: Dr. Jeffrey D. Hart
Dr. Marc G. Genton

This dissertation is composed of four articles describing inference and visualization of periodic sequences.

In the first article, a nonparametric method is proposed for estimating the period and values of a periodic sequence when the data are evenly spaced in time. The period is estimated by a "leave-out-one-cycle" version of cross-validation (CV) and complements the periodogram, a widely used tool for period estimation. The CV method is computationally simple and implicitly penalizes multiples of the smallest period, leading to a "virtually" consistent estimator.

The second article is the multivariate extension, where we present a CV method of estimating the periods of multiple periodic sequences when data are observed at evenly spaced time points. The basic idea is to borrow information from other correlated sequences to improve estimation of the period of interest. We show that the asymptotic behavior of the bivariate CV is the same as the CV for one sequence, however, for finite samples, the better the periods of the other correlated sequences are estimated, the more substantial improvements can be obtained.

The third article proposes an informative exploratory tool, the functional boxplot, for visualizing functional data, as well as its generalization, the enhanced functional boxplot. Based on the center outwards ordering induced by band depth for

functional data, the descriptive statistics of a functional boxplot are: the envelope of the 50% central region, the median curve and the maximum non-outlying envelope. In addition, outliers can be detected by the 1.5 times the 50% central region empirical rule.

The last article proposes a simulation-based method to adjust functional boxplots for correlations when visualizing functional and spatio-temporal data, as well as detecting outliers. We start by investigating the relationship between the spatio-temporal dependence and the 1.5 times the 50% central region empirical outlier detection rule. Then, we propose to simulate observations without outliers based on a robust estimator of the covariance function of the data. We select the constant factor in the functional boxplot to control the probability of correctly detecting no outliers. Finally, we apply the selected factor to the functional boxplot of the original data.

To my family whose love never stopped.

ACKNOWLEDGMENTS

Being a Ph.D. student in the U.S. is one of the most important experiences in my life. I would like to express my deep and sincere gratitude to my advisors, Dr. Marc G. Genton and Dr. Jeffrey D. Hart, for providing me with opportunities and shaping my career, for their guidance, encouragement and support during my graduate studies. Your enthusiasm, inspiration, patience, generosity and great efforts have made my graduate experience a beautiful journey, productive and stimulating. I am proud to have worked with both of you and will always cherish the experience. Without your help, I would never have accomplished as much as I have.

Many thanks are extended to Dr. Kenneth P. Bowman who served on my committee. I would like to thank you for your many helpful comments. Your expertise and advice in meteorology were invaluable. A special thanks to committee member, Dr. Raymond J. Carroll. Thank you for directing me to such an interesting research area at the very beginning and being consistently supportive. Thanks also to Dr. Huiyan Sang for her kind advice from time to time and for being on my defense committee.

I would also like to thank Dr. Ramalingam Saravanan in the Department of Atmospheric Sciences for providing the El Niño/Southern Oscillation (ENSO) data, and Dr. Caspar M. Ammann at the National Center for Atmospheric Research for providing the General Circulation Model (GCM) data, which motivated the methodological research and turned out to be very interesting applications.

Thanks also go to my great department, the wonderful Ph.D. program, the excellent faculty members and all the resources that made everything possible. I would also like to thank the members from the Collaborations in Mathematical Geosciences seminar group who generated many interesting discussions on the topics I presented.

Last but not least, a big thank you to my friends for the good company and to my family for all their love, always nodding and smiling. Without you, it would have been certainly much harder to finish a Ph.D.. Still today, learning to love and to receive love makes me a better person.

Looking back, I am surprised and at the same time very grateful for all I have received throughout these years. It has certainly shaped me into the person I am today and has led me to where I am now. All these years of Ph.D. studies are full of such gifts. I finish with a final silence of gratitude for my life.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

1.1. Period Estimation

Many datasets are periodic or nearly so in time series analysis, and many applications of period estimation methods have been to problems in the physical and environmental sciences. For instance, stars for which brightness changes over time are referred to as variable stars. For many such stars, brightness varies in a periodic, or approximately periodic way. The period lengths, and light-curve shapes, are of significant interest for the reason that if the intensity of light radiating from a star varies in a periodic fashion over time, then there are significant opportunities for accessing information about the star's origins, age and structure (Hall, 2008). Other interesting period estimation problems include the study of the periodicity of sunspot numbers from one of the earliest recorded monthly series (Schuster, 1906) and more modern investigations centering on whether a warming is present in global temperature measurements, such as the periodicity of the El Niño effect, which can have profound effects on local climate. The study of periodicity extends to economics and social sciences, where one may be interested in yearly periodicities in series such as monthly unemployment or monthly birth rates (Shumway and Stoffer, 2000).

Before looking more closely at the particular statistical methods, it is appropriate to mention that two separate, but not necessarily mutually exclusive, approaches to time series analysis exist, commonly identified as the frequency domain approach and the time domain approach.

The time domain approach is generally motivated by the presumption that cor-

---

This dissertation follows the style of *Biometrika*.

relation between adjacent points in time is best explained in terms of a dependence of the current values on past values, and models future value of a time series as a parametric function of the current and past values, i.e., stochastic models. For example, this includes autoregressive (AR), moving average (MA) or autoregressive integrated moving average (ARIMA) models. In the first part of this dissertation, however, we will focus on modeling the periodicity by deterministic models assuming the periodicity is in the mean function, and our method will be nonparametric.

Conversely, the frequency domain approach assumes the primary characteristics of interest in time series analysis relate to systematic sinusoidal variations found naturally in most data. These periodic variations are often caused by biological, physical, or environmental phenomena of interest and expressed as Fourier frequencies being driven by sines and cosines (Shumway and Stoffer, 2000). The partition of the various kinds of periodic variation in a time series is accomplished by evaluating separately the variance associated with each periodicity of interest. In spectral analysis, the periodogram is a measure of the squared correlation of the data with sinusoids oscillating at a frequency of $\omega_j = j/n$, or j cycles in n time points. This is a popular way to discover the periodic components of a time series. The primary justification for the use of frequency domain methods lies in their potential for explaining the behavior of some empirical phenomenon, such as an explanation involving several primary and physically meaningful oscillations. However, for real data examples, usually little is known about the structure of a periodic sequence, and it is thus important to have nonparametric methods of estimating the period.

In our view, no schism divides time domain and frequency domain methodology, because in many cases, the two approaches may produce similar answers and provide more possibilities of interpretation. Hence, in this dissertation, we will show the advantages of our period estimation methodology in the time domain, but compare

the frequency domain methods in a complementary fashion.

For period estimation methods in the time domain, Hall, Reimann and Rice (2000) proposed a nonparametric method for estimating both the period and the amplitude function from noisy observations of a periodic function made at irregularly spaced time points. They were motivated by applications to brightness data on periodic variable stars where the time points of the observations are irregular, due to a star being observed on most nights but not at the same time each night. They also remarked that if observations are made at regularly spaced points in time then identification of the brightness function without a structured model might not be possible. Therefore, they pointed out that the appropriate use of randomness in selecting design points ensures that a consistent period estimator can be constructed. Our point of view in the dissertation is that we evaluate periodicity by nonparametric methods as best we can when the time points are evenly spaced. We propose methodologies to estimate both the period and sequence values for one periodic sequence as well as a multivariate extension where multiple correlated periodic sequences are observed.

1.2. Complex Data Visualization

Data visualization is the graphical representation of information. The step in many statistical analyses involves careful scrutiny of the plotted data. This scrutiny often suggests the method of analysis that will be of use in summarizing the information in the data. The classical boxplot is an example of simple data visualizations that have been used for decades. More complex data visualizations usually involve observations made over time and space, for example functional data or spatio-temporal data. In many statistical experiments, the observations are functions by nature, such as temporal curves or spatial surfaces, where the basic unit of information is the entire

observed function rather than a string of numbers. Ramsay and Silverman (2005) and Ferraty and Vieu (2006) provided various functional data examples and methods for functional data analysis. In the second part of this dissertation, we propose an informative exploratory tool, the functional boxplot analogous to the classical boxplot, for visualizing complex curve or image data as well as detecting possible outliers candidates. Then, considering that the possible correlations in data might affect the outlier detection performance, we propose a bootstrap method to adjust functional boxplots when visualizing functional and spatio-temporal data.

In the period estimation context, to visualize periodic sequences, the functional boxplot can be applied to all the cycle curves where the periodic sequences are broken down into different cycles by the estimated period. Then we can examine which cycles are more representative or which cycles are more outlying with respect to all the possible cycles in a functional boxplot. Moreover, in time series analysis, many datasets have annual cycles, for instance, the monthly sea surface temperatures related to the El Niño effect and measured in degrees Celsius over the east-central tropical Pacific Ocean. The functional boxplot can be used to detect El Niño years if we let each curve represent one year of observed sea surface temperatures. This example will be discussed in detail to illustrate the functional boxplot and the adjusted functional boxplot.

1.3. Overall Structure

The following is the the general structure of the dissertation.

Chapter II and Chapter III are about inference of periodic sequences. In Chapter II, a nonparametric method is proposed for estimating the period and values of a periodic sequence when the data are evenly spaced in time. The period is esti-

mated by a "leave-out-one-cycle" version of cross-validation (CV) and complements the periodogram, a widely used tool for period estimation. The CV method is computationally simple and implicitly penalizes multiples of the smallest period, leading to a "virtually" consistent estimator, which is investigated both theoretically and by simulation. Estimating a period is tantamount to selecting a model, and it is shown that the CV estimator works much better in the period estimation context than it does in other model selection problems. As applications, the CV method is demonstrated on three well-known time series: the sunspots and lynx trapping data, and the El Niño series of sea surface temperatures. Chapter III is a multivariate extension, where we present a method of estimating the periods of multiple periodic sequences when data are observed at evenly spaced time points. The basic idea is to borrow information from other correlated sequences to improve the period estimation of interest.

Chapter IV and Chapter V introduce the visualization tools. Chapter IV proposes an informative exploratory tool, the functional boxplot, for visualizing functional data, as well as its generalization, the enhanced functional boxplot. Based on the center outwards ordering induced by band depth for functional data, the descriptive statistics of a functional boxplot are: the envelope of the 50% central region, the median curve and the maximum non-outlying envelope. In addition, outliers can be detected in a functional boxplot by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. The construction of a functional boxplot is illustrated on a series of sea surface temperatures related to the El Niño phenomenon and its outlier detection performance is explored by simulations. As applications, the functional boxplot and enhanced functional boxplot are demonstrated on children growth data and spatio-temporal U.S. precipitation data for nine climatic regions, respectively. Chapter V proposes a simulation-based method to adjust functional boxplots for correlations when visualizing functional and spatio-temporal data,

as well as detecting outliers. We start by investigating the relationship between the spatio-temporal dependence and the 1.5 times the 50% central region empirical outlier detection rule. Then, we propose to simulate observations without outliers based on a robust estimator of the covariance function of the data. We select the constant factor in the functional boxplot to control the probability of correctly detecting no outliers. Finally, we apply the selected factor to the functional boxplot of the original data. As applications, the factor selection procedure and the adjusted functional boxplots are demonstrated on sea surface temperatures, spatio-temporal precipitation and General Circulation Model (GCM) data. The outlier detection performance is also compared before and after the factor adjustment.

Summary and possible future extensions are discussed in Chapter VI and all the theoretical proofs are provided in the Appendix.

CHAPTER II

NONPARAMETRIC ESTIMATION OF A PERIODIC SEQUENCE

2.1. Introduction

Sequences that are periodic or nearly so appear in many disciplines, including astronomy, meteorology, environmetrics and economics. When little is known about the structure of such a sequence, it is important to have nonparametric methods of estimating both its values and period. This paper is concerned with these two estimation problems in the situation where one has observations at equally spaced time points. We consider the following model:

$$Y_t = \mu_t + \varepsilon_t, \quad t = 1, \ldots, n, \tag{2.1}$$

where $Y_1, \ldots, Y_n$ are the observations, $\mu_1, \ldots, \mu_n$ are unknown constants, and the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent, identically distributed and mean-zero random variables. It is assumed that $\mu_t = \mu_{t+mp}$ for $t = 1, \ldots, p$, $m = 1, 2, \ldots$, and some integer $p \geq 2$. (It is implicit that p is the smallest such integer.) Of interest is estimating the period p and the sequence values $\mu_1, \ldots, \mu_p$. Our approach to the problem will be nonparametric. Even though p is finite, we assume no upper bound on it, and hence there is no bound on how many sequence values have to be estimated.

One motivation for our methodology is the study of periodicity of El Niño effects, which are defined as sustained increases of at least 0.5 °C in average sea surface temperatures over the east-central tropical Pacific Ocean. Such occurrences that last less than five months are classified as El Niño conditions. An anomaly persisting for five months or longer is called an El Niño episode. Typically, this happens at irregular

intervals of 2-7 years and lasts nine months to two years; see Torrence and Webster (1999) and references therein. El Niño is associated with floods, droughts and other weather disturbances in many regions of the world, particularly those bordering the Pacific Ocean. We will apply our proposed methodology to a series of sea surface temperatures in order to estimate the El Niño period.

Our method of estimating the period is based on cross-validation (CV). A candidate period q is evaluated by first "stacking," at the same time point, all data which are separated in time by a multiple of q, and then computing a "leave-out-one-cycle" version of the variance for each of the q stacks of data. This variance will tend to be smallest when q equals p. A sequence of period p is also periodic of period mp for $m = 1, 2, \ldots$, and hence a cross-validatory assessment of variance is required to avoid overfitting, i.e., overestimating the period. It is shown that this method of estimating p is asymptotically equivalent to Akaike's information criterion (AIC, Akaike 1973) when the errors in (2.1) are assumed to be i.i.d. Gaussian.

We show that when p is sufficiently large, our period estimator $\hat{p}$ is *virtually* consistent, in the sense that $\lim_{n\to\infty} P(\hat{p} = p)$ increases to 1 as p increases. When p is 16, for example, $\lim_{n\to\infty} P(\hat{p} = p) \approx 0.990$. These results are surprisingly good in comparison to a variety of model selection problems. Consider the canonical scenario where the true model is of finite dimension k, and one selects a model from a collection of nested models (of which the true model is a member). Here, the asymptotic distribution of AIC and cross-validation estimators of k are typically independent of the value of k. Furthermore, there is usually a substantial probability that these methods will *overestimate* k. For example, when using AIC to select the order of a finite order autoregression, Shibata (1976) shows that the asymptotic probability that AIC overestimates the true order is about 0.29. The fact that p is discrete makes the virtual consistency result remarkably strong. To wit, for all n sufficiently large,

the probability that our period estimator is exactly equal to the truth is at least 0.99 for each $p \geq 16$.

Recent articles on nonparametric methods for estimating periodic functions include Hall, Reimann and Rice (2000), Hall and Yin (2003), Hall and Li (2006), Genton and Hall (2007) and Hall (2008), the last of which is an excellent review of the subject. The assumption in these papers is that the underlying function, f, of interest is defined at all the reals and has period that could be any positive number. As pointed out by Hall (2008), an appropriate *random* spacing of time points is needed in such problems in order to ensure consistent estimation of the period of f. It is important to appreciate that the problem addressed in the current paper is different from that just discussed. We are content to estimate the period of the sequence $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots\}$ even in cases where $\boldsymbol{\mu}$ may be a sampling of an underlying function f. In such cases, f need not be periodic even though $\boldsymbol{\mu}$ is, or if f *is* periodic, then its period need not be the same as that of $\boldsymbol{\mu}$. For equally spaced data, Quinn and Thomson (1991) proposed a periodogram-based method for period estimation which is discussed in Section 2.2.4.

The rest of this chapter is organized as follows. Section 2.2 describes the cross-validation method of estimating the period and sequence values, while Section 2.3 discusses asymptotic properties of the method. Simulations motivated by real data applications are reported in Section 2.4, and Section 2.5 illustrates applications of our methods to well-known time series, including the El Niño series of sea surface temperatures. Concluding remarks are provided in Section 2.6, including a discussion of how our period estimator performs when the underlying periodic function has domain $\mathbb{R}^+$. The proof of virtual consistency is provided in Appendix A.

2.2. Methodology

Suppose we observe a time series $\{Y_t : t = 1, \ldots, n\}$ from the model (2.1), where the sequence $\boldsymbol{\mu}$ is periodic with (smallest) period p and $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with zero means and finite variances σ^2. We propose a methodology for estimating p in Section 2.2.1 and discuss some related model selection criteria in Section 2.2.2. In Section 2.2.3 we propose methods for estimating the sequence values $\mu_1, \ldots, \mu_p$ within one period.

2.2.1. Cross-validation method for estimating an integer period

Let q be a candidate integer period with $2 \le q \le M_n$, where M_n is of smaller order than $\sqrt{n}$. For each $i = 1, \ldots, q$, we construct an estimator of μ_i by stacking all data that are separated in time by a multiple of q. So, at time point i we have data $Y_i, Y_{i+q}, \ldots, Y_{i+qk_{q,i}}$, where $k_{q,i}$ is the largest integer such that $i + qk_{q,i} \le n$. For each relevant q, i and j, define $Y_{qij} = Y_{i+(j-1)q}$ and $\mu_{qij} = \mu_{i+(j-1)q}$. Let

$$\mathrm{CV}(q) = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} (Y_{qij} - \bar{Y}_{qi}^{j})^2, \tag{2.2}$$

where $\bar{Y}_{qi}^{j}$ is the average of $Y_{qi\ell}$, $\ell = 1, \ldots, k_{q,i}$, excluding Y_{qij}. We define a period estimator $\hat{p}$ to be the minimizer of $\mathrm{CV}(q)$ for $2 \le q \le M_n$.

The version of cross-validation used in (2.2) is more closely related to that of Hart and Wehrly (1993) in the setting of functional data than to the version used in smoothing independent data. The method of Hart and Wehrly (1993) leaves out one curve in a sample of curves and then predicts the deleted curve by using all the others. Analogously in our setting, all the data in one cycle (corresponding to a putative cycle length) are deleted, and then data from other cycles are used to predict the omitted cycle. One may regard different cycles as independent copies of a multivariate random

variable in the same way that functional data are independent copies of a curve.

The CV method may be motivated by taking the expectation of (2.2). Let $\varepsilon_{qij} = Y_{qij} - \mu_{qij}$ and $\bar{\varepsilon}^j_{qi} = \bar{Y}^j_{qi} - \bar{\mu}^j_{qi}$, where $\bar{\mu}^j_{qi}$ is the average of $\mu_{qi\ell}$, $\ell = 1, \ldots, k_{q,i}$, excluding μ_{qij}. Then (2.2) may be written as

$$\mathrm{CV}(q) = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} (\varepsilon_{qij} - \bar{\varepsilon}^j_{qi} + \mu_{qij} - \bar{\mu}^j_{qi})^2.$$

Now, there are at most two distinct values of $k_{q,i}$ and these values differ by only 1. So,

$$\begin{aligned} E\{\mathrm{CV}(q)\} &\approx E\{(\varepsilon_{q11} - \bar{\varepsilon}^1_{q1})^2\} + \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} (\mu_{qij} - \bar{\mu}^j_{qi})^2 \\ &\approx \sigma^2 \left(1 + \frac{1}{k_{q,1} - 1}\right) + \frac{k^2_{q,1}}{(k_{q,1} - 1)^2} \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} (\mu_{qij} - \bar{\mu}_{qi})^2 \\ &= \sigma^2 \left(1 + \frac{1}{k_{q,1} - 1}\right) + \frac{k^2_{q,1}}{(k_{q,1} - 1)^2} C_q \\ &= E_q, \end{aligned}$$

where $\bar{\mu}_{qi}$ is the average of $\mu_{qi\ell}$, $\ell = 1, \ldots, k_{q,i}$.

A motivation for our cross-validation method stems from the following facts:

F1. The term C_q is 0 if and only if q is a multiple of p, as shown in Appendix A.

F2. By fact F1, $\min_{2 \le q < p} E_q > E_p$ for all n sufficiently large.

F3. Fact F1 and the definition of $k_{q,i}$ imply that $\min_{p < q \le M_n} E_q > E_p$ for all n sufficiently large.

More succinctly, we may say that, over any bounded set of q-values containing p, the minimizer of $E\{\mathrm{CV}(q)\}$ is p for all sufficiently large n.

The *necessity* of using cross-validation follows upon considering what happens

when $\bar{Y}_{qi}^{j}$ in (2.2) is replaced by $\bar{Y}_{qi}$, the average of $Y_{q1}, \ldots, Y_{qk_{q,i}}$. Call the resulting criterion $V(q)$. Then for $q = kp$ and $k = 1, 2, \ldots$,

$$E\{V(q)\} \approx E\{(\varepsilon_{q11} - \bar{\varepsilon}_{q1})^2\} = \sigma^2 \left(1 - \frac{1}{k_{q,i}}\right) \approx \sigma^2 \left(1 - \frac{kp}{n}\right).$$

The rightmost quantity immediately above is *decreasing* in k, and so the minimizer of $V(q)$ will tend to be a large multiple of p.

Of course, the considerations above do not prove that CV yields a good estimator of p. That question will be addressed in Section 2.3.

2.2.2. Model selection criteria for period estimation

Cross-validation is typically used as a model selection tool. Our use of CV may appear to be different, but in fact each candidate period q corresponds to a model for the sequence $\boldsymbol{\mu}$ consisting of the q parameters $\mu_1, \ldots, \mu_q$. With this perspective it is thus natural to consider model selection criteria other than CV in estimating q.

If we assume that the errors in model (2.1) are i.i.d. Gaussian, then Akaike's information criterion has the form

$$\mathrm{AIC}(q) = n\log\hat{\sigma}_q^2 + 2(q+1), \quad q = 2, \ldots, M_n,$$

where $\hat{\sigma}_q^2 = V(q)$, as defined in Section 2.2.1. Minimizing AIC(q) with respect to q provides an estimator of p. It is not difficult to argue that under appropriate conditions on M_n, the estimator of p obtained from AIC is asymptotically equivalent to the CV estimator of Section 2.2.1.

The usual variations of AIC are also possible. These have the general form

$$C(q) = n\log\hat{\sigma}_q^2 + c_n(q+1), \quad q = 2, \ldots, M_n, \tag{2.3}$$

for some (specified) penalty c_n. Bayes information criterion (Schwarz 1978), or BIC,

corresponds to $c_n = \log n$. Another possibility is a Hannan-Quinn-type criterion with $c_n = \text{loglog}n$ (Hannan and Quinn 1979). Both these methods have $c_n > 2$ for all n large enough, and hence have a smaller probability of overestimating the true model dimension than does AIC. In fact, BIC and the Hannan-Quinn criterion produce consistent estimators of the true model dimension, whereas CV and AIC do not. This is true in the setting of the current paper as well as in many scenarios where a sequence of *nested* models is under consideration.

It will be shown in Section 3 that although CV (and hence AIC) do not produce consistent estimators of p, their asymptotic probability of selecting p is very close to 1 for p larger than 15. We call this last property *virtual* consistency. The virtual consistency of CV/AIC is somewhat surprising in comparison to familiar results associated with CV and AIC. For example, if the true model is of finite dimension and AIC chooses amongst nested models, then its asymptotic probability of selecting the correct model is 0.71, independent of the model dimension (Shibata 1976 and Woodroofe 1982). In spite of its lack of consistency, AIC is still preferred to BIC by some practitioners, since the latter criterion has a higher likelihood of underestimating the model dimension. In our setting this preference is even more justified since CV/AIC have a much higher than usual probability of selecting the correct model.

The CV method is desirable because of its virtual consistency, but if a formally consistent criterion is desired, then one of the form (2.3) would be an option. Among them, one that we find appealing has $c_n = \max(2, \text{loglog}n)$, which is a compromise between AIC and the Hannan-Quinn-type criterion. On the one hand it yields a consistent estimator of p, but on the other it matches the definition of AIC for all $n < 1619$, since then $\text{loglog}n \leq 2$.

2.2.3. Estimating sequence values

Given that the period is q, we desire estimates for the sequence values $\mu_1, \ldots, \mu_q$. If the errors are assumed to be i.i.d. Gaussian, then it is straightforward to verify that the maximum likelihood estimates of these parameters are the means $\bar{Y}_{q1}, \ldots, \bar{Y}_{qq}$, as defined in Section 2.2.1. For the moment, we set aside the effect that estimation of p has on estimates of the sequence values, and assume that $q = p$. Obviously, whenever the errors have two finite moments, $\bar{Y}_{p1}, \ldots, \bar{Y}_{pp}$ are consistent for $\mu_1, \ldots, \mu_p$, since as n tends to ∞, the number of observed full cycles, $k_{p,i}$, also tends to ∞.

James-Stein theory (Stein 1964) implies that the estimator $\bar{\boldsymbol{Y}} = (\bar{Y}_{p1}, \ldots, \bar{Y}_{pp})$ is inadmissible when $p \geq 3$. A better estimator of a normal mean vector may be obtained by shrinking the sample mean vector towards a specified point. In our setting where the data are time ordered, a natural form of shrinkage is to smooth means that are close to each other in time. More precisely, we may estimate μ_i by

$$\hat{\mu}_i = \sum_{t=1}^{q} \bar{Y}_{pt} K_h(|t - i|), \tag{2.4}$$

where $\sum_{t=1}^{q} K_h(|t - i|) = 1$, K_h decreases monotonically to 0 on $[0, \infty)$ and h is a smoothing parameter that dictates how quickly K_h goes to 0. When the sequence $\boldsymbol{\mu}$ is "smooth," in the sense that $\sum_{t=1}^{p-1}(\mu_{t+1} - \mu_t)^2$ is sufficiently small, then the kernel-type smoother (2.4) can have smaller mean squared error than $\bar{Y}_{pi}$.

Asymptotically, the use of kernel smoothing cannot be expected to yield a large improvement over the simple mean $\bar{Y}_{pi}$. This is because an estimator of the form (2.4) must eventually collapse to $\bar{Y}_{pi}$ in order to be consistent. This entails that an optimal estimator of the form (2.4) will have mean squared error asymptotic to $p\sigma^2/n$, the same as that of $\bar{Y}_{pi}$. The situation where a substantial improvement could be obtained by smoothing is when n is small enough that p/n is fairly large. In this case

the variance of $\bar{Y}_{pi}$ can be large, and so smoothing, while introducing bias, can be beneficial if it reduces variance substantially.

Obviously, maximum likelihood estimates of sequence values can be profoundly affected by misspecification of the value of p. Fortunately, our CV method is such that, asymptotically, it only chooses p or a multiple of p, and if p is more than 15, the probability of choosing a value larger than p is extremely small. Let $\hat{p}$ be the estimate of p obtained with the CV method. A useful check on this estimate is to plot $\hat{\mu}_1, \ldots, \hat{\mu}_{\hat{p}}$ against time and check for evidence of more than one cycle. If, for example, a pattern seems to repeat itself twice in the sequence of $\hat{\mu}_i$s, then there is evidence that $\hat{p} = 2p$.

2.2.4. Cross-validation method versus the periodogram

Quinn and Thomson (1991) (QT) proposed a periodogram-based method for estimating the period of a signal when the data are observed at the time points $1, 2, \ldots$. Their method is more general than ours in two ways. Firstly, their method can sometimes consistently estimate a noninteger period, and secondly, they allow the errors to come from a covariance stationary process.

QT assume that observations $Y_1, \ldots, Y_n$ follow a model of the form

$$Y_j = r(j) + \epsilon_j, \quad j = 1, \ldots, n,$$

where r is defined for all $t > 0$ and is periodic with arbitrary period p. They model r as the Fourier series

$$r(t) = \mu + \sum_{j=1}^{m} \rho_j \cos\left(\frac{2\pi t j}{p} - \phi_j\right), \quad t > 0.$$

When the errors are white noise (as in our setting), the QT estimate of p is asymptotically equivalent to the value of q that maximizes $\mathcal{C}(q) = \sum_{j=1}^{m} \hat{f}(j/q)$ over $q \geq 2m$,

where $\hat{f}(\omega)$, $0 \leq \omega \leq 1/2$, is the periodogram. There are two potential problems with this approach. First of all the method requires that m be smaller than $p/2$, and since p is unknown it is unclear that an appropriate choice for m will satisfy this requirement. A second problem is that even if one is certain that $p \geq p_0$ and that $m \leq p_0/2$ is satisfactory, it is still important to make a reasonable choice of m. In particular, too small an m can lead to inconsistent estimation of p. Suppose, for example, that m is taken to be 1 but in fact the Fourier series has more than one term. Then the maximizer of $\mathcal{C}(q)$ will be inconsistent for p unless the largest of $|\rho_1|, \ldots, |\rho_m|$ happens to be $|\rho_1|$.

Concerning the error series, we note that our method is not necessarily unsound when the errors are dependent. While it is beyond the scope of the current paper to theoretically investigate the effect of dependent errors, we will use simulation in Sections 4.2 and 5.3 to show that the cross-validation method can be quite robust to serial correlation that is negligible at lags greater than the period.

The method of QT is a very useful tool for investigating periodicity, but it does have its shortcomings. We feel that our cross-validation method can be a valuable complement to that of QT, especially when there are doubts about the assumptions in the latter method.

2.3. Asymptotic Properties of the CV Method

We begin with a theorem that describes the asymptotic behavior of the CV period estimator.

Theorem 1 *Suppose that model* (2.1) *holds with* $\varepsilon_1, \ldots, \varepsilon_n$ *independent and identically distributed* $N(0, \sigma^2)$ *random variables. The sequence* $\boldsymbol{\mu}$ *is assumed to be periodic with smallest period* p*, which is a positive integer. Let* $\{M_n : n = 1, 2 \ldots\}$ *be a*

sequence of positive integers that tends to infinity and is $o(\sqrt{n})$*, and define*

$$\hat{p} = \text{argmin}_{q \in \{1,\ldots,M_n\}} CV(q).$$

Let $\{Z_{m,i} : m = 1, 2, \ldots; i = 1, \ldots, pm\}$ *be a collection of random variables such that each* $Z_{m,i} \sim N(0,1)$ *and any finite subset has a multivariate normal distribution. Furthermore,*

C1. For each m*, the random variables* $Z_{m,1}, \ldots, Z_{m,pm}$ *are mutually independent.*

C2. For each $r > m$*,* $Z_{m,i}$ *and* $Z_{r,j}$ *are independent unless* $i - j = p(rs - m\ell)$ *for some pair of nonegative integers* ℓ *and* s*, in which case*

$$\text{Corr}(Z_{m,i}, Z_{r,j}) = \frac{1}{\sqrt{mr}}.$$

Define $S_m = \sum_{i=1}^{pm} Z_{m,i}^2$ *for* $m = 1, 2, \ldots$*. Then if* $\mathcal{S}_n$ *is the set* $\{1, 2, \ldots, M_n\}$ *excluding multiples of* p*,*

$$\lim_{n\to\infty} P(\hat{p} \in \mathcal{S}_n) = 0,$$

and

$$\lim_{n\to\infty} P(\hat{p} = jp) = P\left(\bigcap_{m=1}^{\infty} \{S_j - S_m + 2p(m-j) \geq 0\}\right), \quad j = 1, 2, \ldots.$$

Defining $\tau(p) = \lim_{n\to\infty} P(\hat{p} = p)$, we have

$$\tau(p) \geq 1 - P\left(S_1 - S_2 + 2p < 0\right) - \sum_{m=3}^{\infty} P(S_m > 2p(m-1)). \tag{2.5}$$

The random variable $(S_1 - S_2)/p$ converges in probability to -1 as $p \to \infty$ and hence $P(S_1 - S_2 + 2p < 0)$ tends to 0 as $p \to \infty$. By applying Bernstein's inequality to each term of the infinite series on the right-hand side of (2.5), that series may be bounded by a quantity that tends to 0 as $p \to \infty$. These facts show that $\tau(p)$ tends to 1 as $p \to \infty$, which proves the virtual consistency of $\hat{p}$.

For small values of p, simulation of the asymptotic distribution of $\hat{p}$ was conducted to approximate $P(\hat{p} = p)$. The results are shown in Table 1 and were obtained using $n = 5000$, $M_n = 20p$ and 5000 replications for each p.

Table 1: Asymptotic probability of choosing the correct period for small p.

p	$\hat{\tau}(p)$	p	$\hat{\tau}(p)$
1	0.489	9	0.957
2	0.694	10	0.968
3	0.791	11	0.971
4	0.854	12	0.977
5	0.892	13	0.981
6	0.908	14	0.983
7	0.933	15	0.988
8	0.954	16	0.990

The asymptotic probability of choosing the correct period increases quickly as p increases. The probability is about 0.96 for p as small as 9 and increases to about 0.99 when p is 15.

Using test inversion and the result of Theorem 1, one may obtain asymptotic $(1-\alpha)100\%$ confidence sets for p. A little thought makes it clear that such a set is a subset of $\{\hat{p}, \hat{p}/k_1, \ldots, \hat{p}/k_{j(p)-1}, 1\}$, where $2 \leq k_1 < k_2 < \cdots < k_{j(p)} = \hat{p}$ are all the divisors of $\hat{p}$. This entails that for prime $\hat{p}$, the confidence set contains at most $\hat{p}$ and 1. Let p_0 be the observed value of $\hat{p}$ and p_0/k be a candidate for the confidence set. Then p_0/k is included in this set if and only if $P(\hat{p} \geq p_0 | p = p_0/k) > \alpha$, where P denotes probability for the limiting distribution of $\hat{p}$. Knowledge of the asymptotic

distribution of $\hat{p}$ allows one to determine conditions under which the large sample confidence set contains only $\hat{p}$. For $\alpha = 0.05$, simulation indicates that this occurs when $\hat{p}$ is at least 21.

One may justifiably be skeptical of using large sample theory to obtain a confidence set when n is not "large." Alternatively, a parametric bootstrap could be used to approximate confidence sets in small samples. As before this can be done using test inversion. To determine if a given value of p, say $\tilde{p}$, is included in the confidence set, one may approximate the probability distribution of $\hat{p}$ on the assumption that $p = \tilde{p}$. This can be done as follows. Let $\hat{\mu}_1, \ldots, \hat{\mu}_{\tilde{p}}$ and $\hat{\sigma}^2$ be the Gaussian maximum likelihood estimates of $\mu_1, \ldots, \mu_{\tilde{p}}$ and σ^2 on the assumption that $\tilde{p}$ is the true period. Then one may generate many samples of size n from a version of model (2.1) that has period $\tilde{p}$, $\mu_1, \ldots, \mu_{\tilde{p}}$ equal to $\hat{\mu}_1, \ldots, \hat{\mu}_{\tilde{p}}$ and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. as $N(0, \hat{\sigma}^2)$. From each sample the CV estimate, $\hat{p}^*$, of period is obtained, and the empirical distribution of all these estimates provides an approximation to the required distribution.

2.4. Simulation Studies

2.4.1. Virtual consistency

To study the behavior of the CV period estimator, we performed simulations with an intermediate value of the period, i.e., $p = 43$, which is our estimate of the El Niño period. First, we consider model (2.1) with $Y_t = a \sin(2\pi t/43) + \varepsilon_t$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. normal with mean 0 and variance σ^2. In this case $\mu_t = a \sin(2\pi t/43)$ and $p = 43$ is the true period of the sequence $\boldsymbol{\mu}$. Obviously, strong signals and low error levels would lead to a more efficient estimate, and indeed what matters is the ratio σ/a. Thus, in our simulation studies, we fix $a = 1$ and consider $\sigma = 0.5, 1, 1.5$. For all cases, the set over which the objective function $\mathrm{CV}(q)$ was searched was taken to

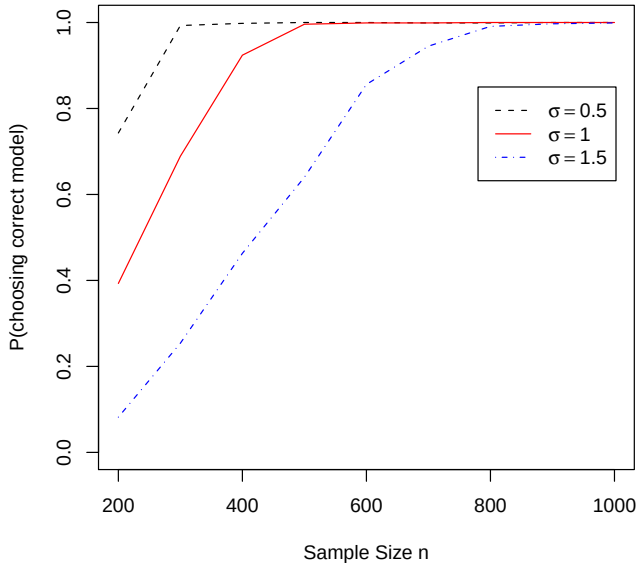

Figure 1. $P(\hat{p} = p)$ for the CV method at different error levels.

be $\{12, 13, \ldots, 96\}$, and the number of replications of each setting is 1000.

Figure 1 shows how the probabilities of choosing $p = 43$ increase as the sample size n increases for each $\sigma = 0.5, 1, 1.5$. It is clear that the convergence is slower for larger errors. For $\sigma = 1$, $\hat{p}$ is between 41 and 44 approximately 86% of the time at $n = 200$ and about 99% of the time at $n = 300$. Averages of CV curves are shown in Figure 2. Here we see what is typical of individual curves, namely they tend to have local minima at or near multiples of p. For example, at $\sigma = 1$, the plots of $E\{\mathrm{CV}(q)\}$ (left panels of Figure 2) are minimized at the true $p = 43$ and have a local minimum at its multiple 86.

Comparing the top and bottom panels of Figure 2 shows how much closer the average CV curves agree with the theoretical expectation when the sample size increases from 200 to 1000. Also it is seen that at $n = 1000$ the value of the CV curve

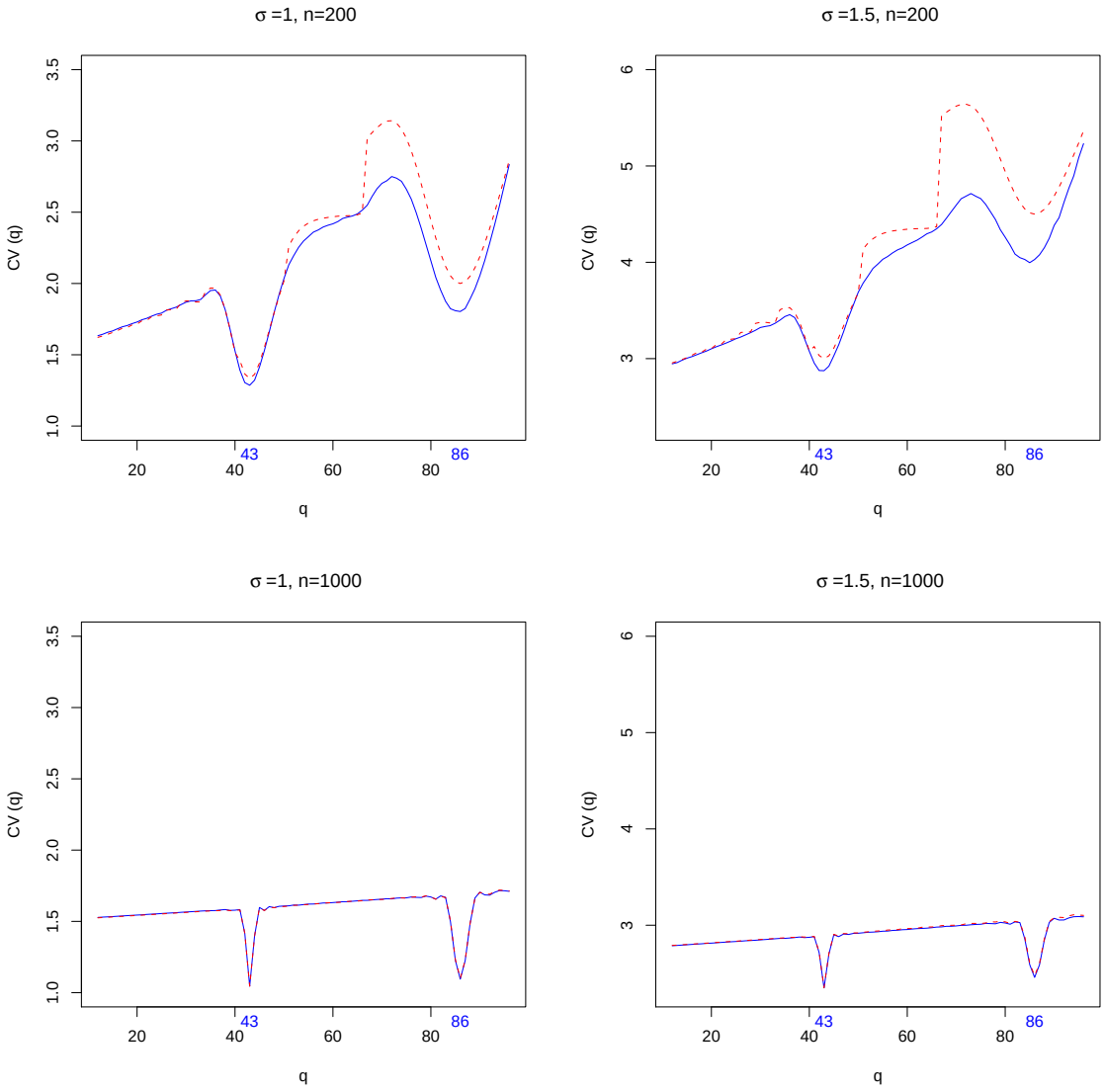

Figure 2. Objective function plot for period candidates q from 12 to 96. The solid blue line denotes theoretical expectation of CV(q) and the dotted red line is the average of all 1000 CV curves from a given simulation.

at its minimum is very close to σ^2.

2.4.2. Robustness of the CV method

Model (2.1) assumes homogeneity of the error terms. In practice, however, many data sets exhibit periodicity but with cycle amplitudes that appear to change randomly. Such behavior can be modeled with heteroscedastic errors. To show the robustness of the CV method for this situation, we simulate data with a periodic mean function and errors whose standard deviations change randomly from one cycle to the next. Specifically, suppose that model (2.1) holds with $\mu_t = \sin(2\pi t/43)$, $n = 43k$ and, conditional on $\sigma_1, \ldots, \sigma_k$, $\varepsilon_1, \ldots, \varepsilon_n$ are independent and normally distributed with 0 means and variances as follows:

$$\mathrm{Var}(\varepsilon_{i+43(j-1)}) = \sigma_j^2, \quad j = 1, \ldots, k, \ i = 1, \ldots, 43.$$

The standard deviations $\sigma_1, \ldots, \sigma_k$ are taken to be i.i.d. Gamma(1,1) with mean 1. Figure 3 shows one simulated example where the data are generated as just described with $k = 10$, i.e. $n = 430$. Simulations at various sample sizes and using our heteroscedastic model were conducted. The results are shown in Figure 3. The proportion of cases in which $\hat{p}$ was equal to p was only slightly less for the heteroscedastic errors than it was for homoscedastic errors with $\sigma = 1$.

Especially for time series data, another interesting question is the robustness of the CV method to serial correlation among the data. We anticipate that the result of our Theorem 1 is not greatly affected by errors that follow an mth order moving average (MA(m)) process as long as $m \leq p$. Presumably the method will also be fairly robust to other types of covariance stationary errors for which the autocorrelation function damps out quickly at lags larger than p. A simulation example shows this to be the case, at least when the errors are first order autoregressive. The right panel

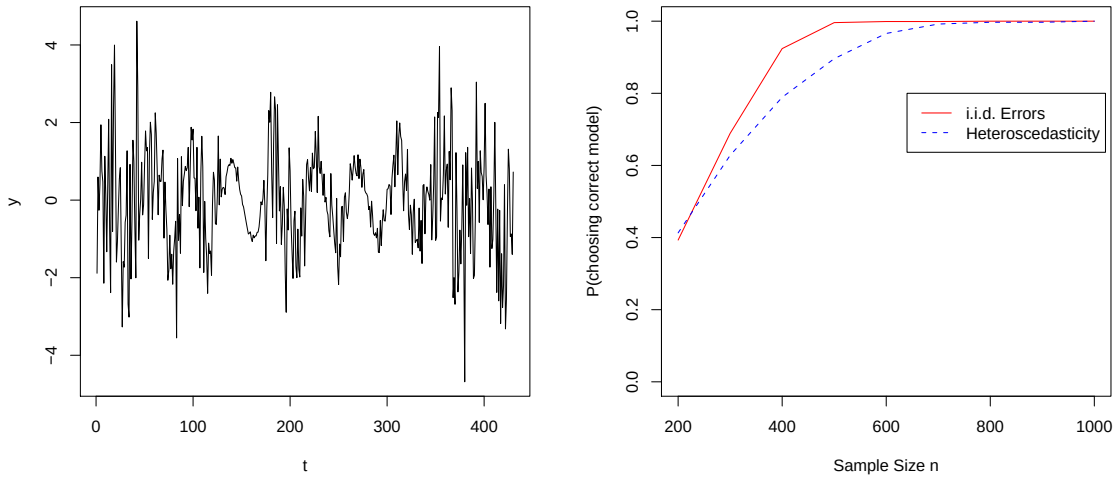

Figure 3. Left panel: an example of heteroscedastic errors with standard deviations $\sigma_1, \ldots, \sigma_k$ taken to be i.i.d. Gamma(1,1) and $p = 43$, $n = 430$. Right panel: virtual consistency of CV method for heteroscedastic errors, $p = 43$.

of Figure 4 gives the proportion of cases with $\hat{p} = p$ when the errors follow a first order autoregressive process with first lag correlation equal to 0.3 and 0.6. While the proportion of correct identifications is less than in the case of i.i.d. errors, the results are nonetheless encouraging.

2.5. Applications

2.5.1. Sunspots and Lynx data

Here we apply our CV method to two classical time series, the sunspots data and the lynx data, both of which are available in `R` (R Development Core Team 2010). The sunspots series consists of mean monthly relative sunspot numbers; see, e.g., Andrews and Herzberg (1985). The data cover the period from 1749 to 1983 for a

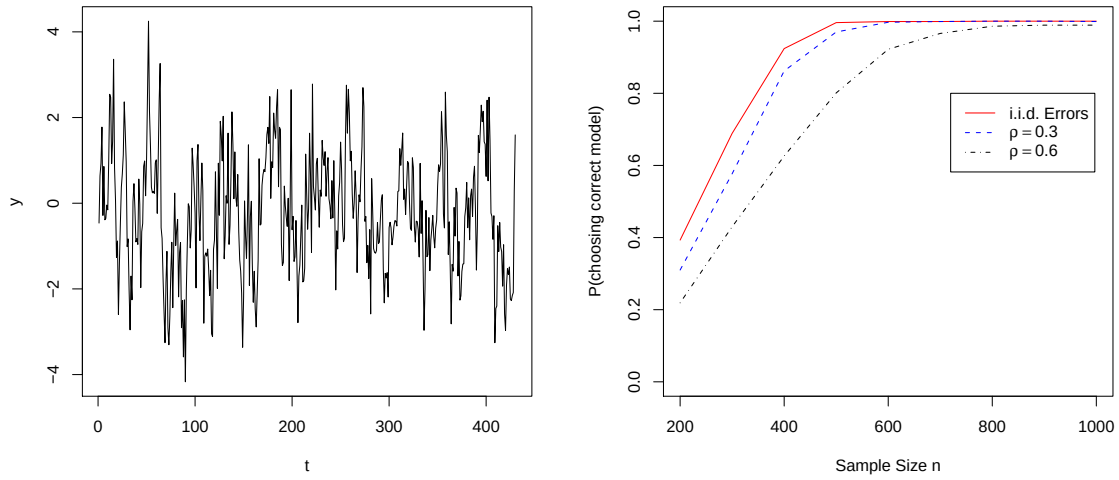

Figure 4. Left panel: an example of autoregressive errors with first lag correlation equal to 0.6, $p = 43$, $n = 430$ and $\sigma = 1$. Right panel: virtual consistency of the CV method for $p = 43$ and errors that follow a first order autoregressive process with $\sigma = 1$.

total of 2820 observations. Sunspots are temporary phenomena on the surface of the sun that appear visibly as dark spots compared to surrounding regions. The number of sunspots peaks periodically, as shown in the left panel of Figure 5. By minimizing the CV objective function (right panel of Figure 5), we get a period estimate of $\hat{p} = 133$ months. This is in close agreement with the "accepted" period estimate of 11 years for the sunspot series. See, for example, the frequency domain analysis of Brockwell and Davis (1991). With such a large $\hat{p}$, the asymptotic 95% confidence set contains only $\hat{p}$, 133 months, as described in Section 2.3.

Another well-known time series consists of the annual number of lynx trappings in Canada from 1821 to 1934. These data have been analyzed by a number of researchers, including Campbell and Walker (1977). The 114 consecutive observations are plotted

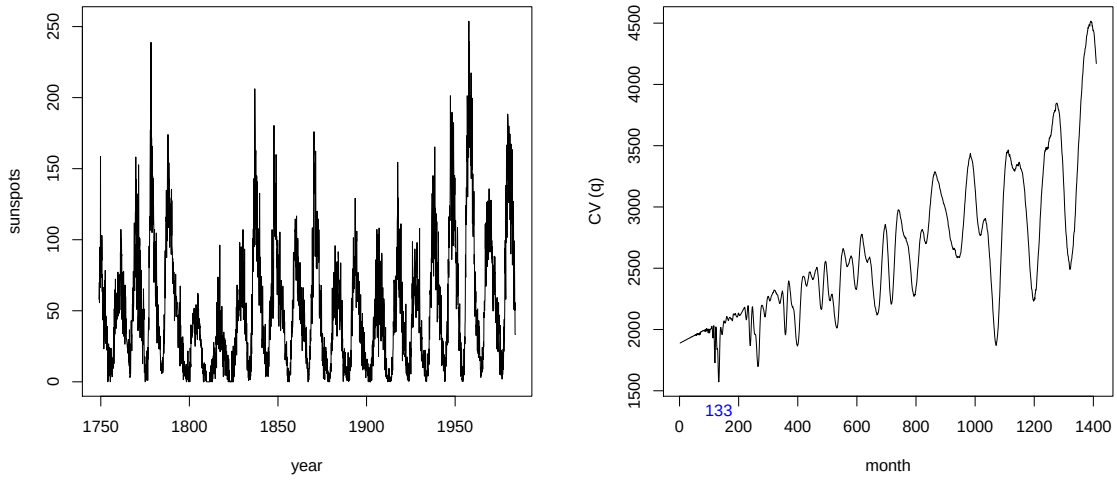

Figure 5. Sunspots data: the left panel is a plot of sunspots numbers from 1749 to 1983 and the right panel is a plot of the CV curve.

in the left panel of Figure 6 and the CV function is shown in the right. The minimizer of the CV function occurs at 38 years, a cycle length that contains four peaks in the time series plot.

The next smallest local minimum is at 19 years, a cycle length that is consistent with a period of 9.5 years. Interestingly, the periodogram for the lynx data is maximized at 9.5 years, which, along with the sinusoidal appearance of the data, is highly suggestive of a 9.5 year period. This is a good example of how our CV method and the periodogram complement each other in cases where the period is not necessarily a multiple of the time interval between data points.

How can one reconcile the disparate estimates of 9.5 and 38 years? In a sense the "correct" estimate depends on how one defines periodicity. If the period is defined as the time between two peaks, then the 9.5-year estimate seems better. However, as can be seen in the time series plot, there is a pattern of one high peak followed

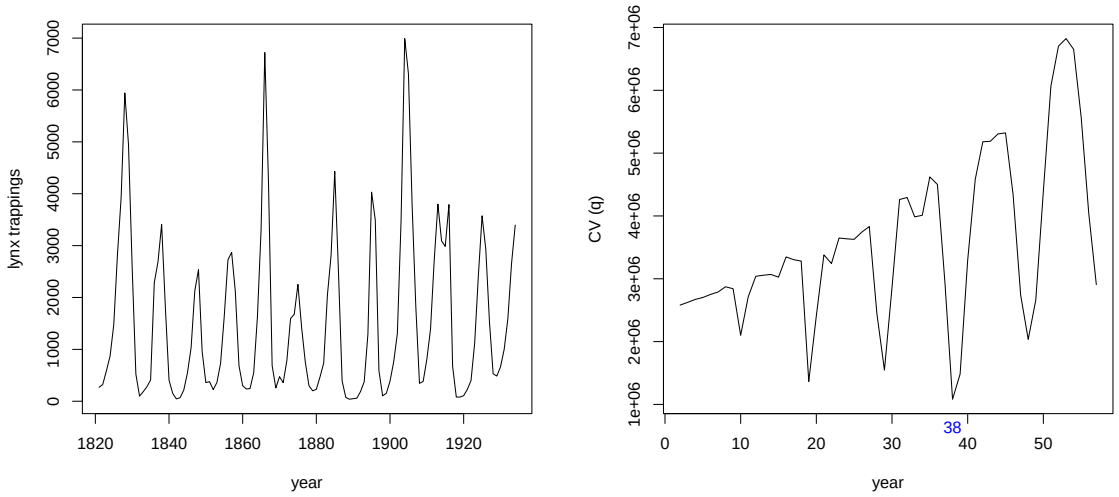

Figure 6. Lynx data: the left panel is a plot of lynx trappings from 1821 to 1934 and the right panel is a plot of the CV curve.

by three lower ones that repeats itself for (almost) three cycles. Therefore, a period estimate of 38 years containing four peaks is perhaps more reasonable according to the strict definition of periodicity. Certainly, more observations would be needed to make the 38 year cycle more convincing. In our opinion it is a strength of the CV method that, in conjunction with the periodogram, it identifies both 9.5 and 38 as possible estimates, since those periods are consistent with what a data plot suggests are plausible possibilities.

2.5.2. El Niño effect

El Niño is a phenomenon associated with warmer than normal sea surface temperatures (SST). It occasionally forms across much of the tropical eastern and central Pacific. The times between successive El Niño events are irregular, but the events tend to recur every 2 to 7 years, as shown by Torrence and Webster (1999). Niño3 SST is an oceanic component index which is one of the measures of variability in

the ENSO-Monsoon system. Estimating the period of the El Niño effect by analyzing Niño3 SST data is a very interesting application in climate science. The data we have consist of the area-average SST over the eastern equatorial Pacific, monthly from January 1950 to February 2009, as seen in the left panel of Figure 7.

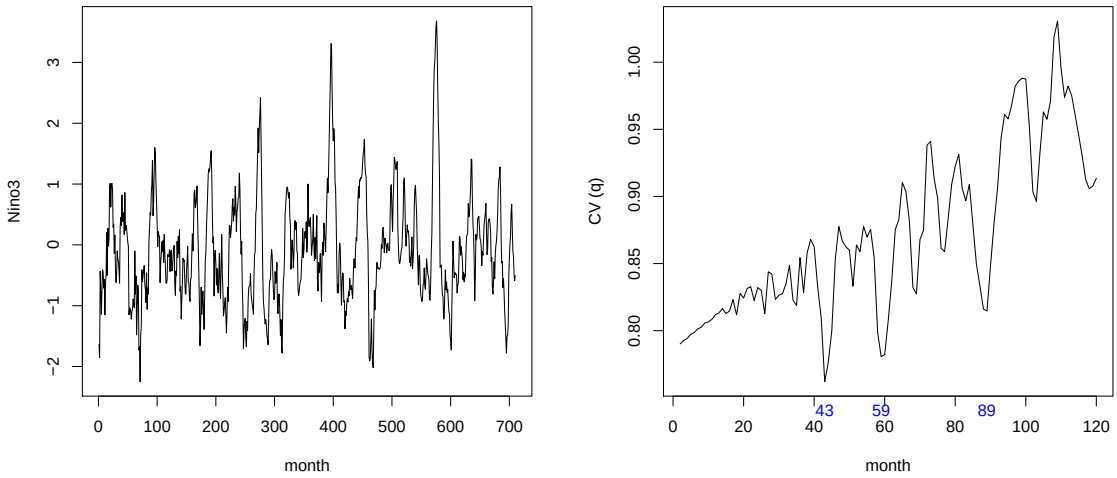

Figure 7. El Niño data: the left panel is a plot of monthly Niño3 sea surface temperatures from 1950 to 2009 and the right panel is a plot of the CV curve.

Our cross-validation method yields a period estimate of $\hat{p} = 43$ months; see the right panel of Figure 7 for the CV curve. The 43 estimated means plot is shown in Figure 8. Notice that the CV curve is also locally minimized at 59 and 89 months. Interestingly, the periodogram is maximized at period 60, and the next largest local maximum is at 42.4 months, both of which are consistent with our CV curve. The other local minimizer, 89 months, is not surprising as it is close to the multiple of 43 months.

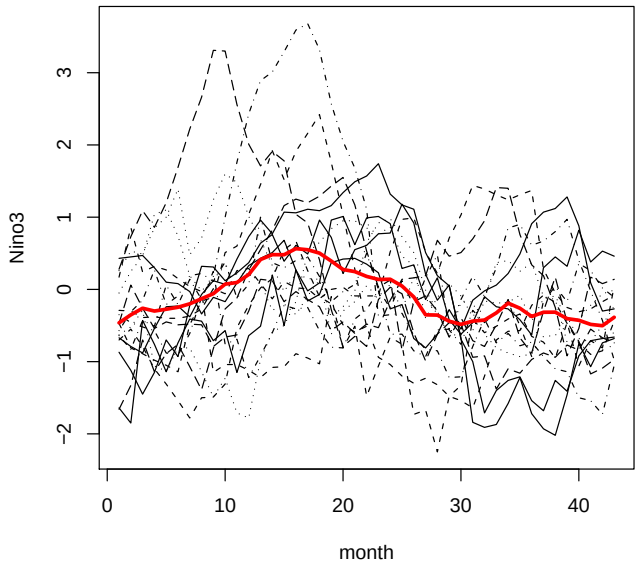

Figure 8. A plot of the 16 1/2 cycles of data (corresponding to a period of 43 months) superimposed on each other. The 43 estimated means are connected by the red line.

2.5.3. Correlation in errors

The errors for any one of the sunspots, lynx or El Niño series are likely to be serially correlated. We thus investigated the effect of correlation by using a bootstrap method. For each series, residuals were computed after obtaining estimates $\hat{p}$ and $\hat{\boldsymbol{\mu}}$ of the period and mean sequence, respectively. ARMA models were fitted to these residuals and AIC was used to choose the ARMA order. Then 1000 independent data sets, each of the same size as the original series, were generated from a model in which the true period and mean sequence were equal to $\hat{p}$ and $\hat{\boldsymbol{\mu}}$ and the error series followed an ARMA process with order and parameters estimated from the original series. The period was estimated by the CV method for each of the 1000 data sets. The results are summarized in Table 2.

In the sunspots simulation, in the twelve cases where the estimated period was

not 133, it was $2 \times 133 = 266$. The lower probability for the El Niño series is due to the fact that its error standard deviation of 0.327 is relatively large. When this standard deviation was lowered to 0.10 and the simulation repeated, the period estimate was 43 in 933 cases and $2 \times 43 = 86$ in the other 67.

These results provide more evidence that the CV method is quite robust to correlation.

Table 2: Estimated probabilities (from 1000 replications) that $\hat{p} = p$ when the errors are correlated as in the real data examples.

Series	Error model	Estimated probability
Sunspots	ARMA(6,4)	0.988
Lynx	ARMA(5,2)	0.999
El Niño	ARMA(5,3)	0.681

2.6. Discussion

2.6.1. Periodic function with domain $\mathbb{R}^+$

In some cases the observed data are a sampling of some periodic function f that is defined on all of, say, $\mathbb{R}^+$. In this case the period of f could be any positive number, and in particular is not necessarily a multiple of the spacing between consecutive time points.

Hall (2008) remarks that in such cases the time points at which data are observed should not be equally spaced, since then "for many values of the period (in particular when the period is a rational multiple of the spacing), consistent estimation is not possible." So, when the period of f can be arbitrary, *design*, i.e., placement, of the

time points at which observations are made, is important. Hall (2008) points out that the appropriate use of randomness in selecting design points ensures that a consistent period estimator can be constructed.

Our point of view in the current paper is that we will try to evaluate periodicity as best as we can when the time points *are* evenly spaced, which is obviously a prevalent situation in time series analysis. Suppose that the observations are taken at time points $1, 2, \ldots$. Then the estimable parameters are the function values $f(1), f(2), \ldots$, which need not be periodic even though f is. Conversely, if the sequence *is* periodic, its period need not be the same as that of f. However, when the period of f is sufficiently large, then practically speaking the sequence $f(1), f(2), \ldots$ will have the same period as f. Furthermore, as explained in Section 2.5.1, the periodogram can be a valuable complementary tool to our time domain methodology in situations where the period of f is not an integer.

If the period of f is a rational number, then the sequence $f(1), f(2), \ldots$ will be periodic with some (smallest) integer period p. In such a case the CV estimator $\hat{p}$ will obviously be estimating p rather than the rational period of f. Nonetheless, there are still other possible ways one can investigate the possibility of a noninteger rational period. One may check to see if the CV criterion has local minima at integers near $\hat{p}/2, \hat{p}/3, \ldots$. For example, if $\hat{p} = 37$, then a rational period of 18.5 is a possibility, and it would be worthwhile to see if the CV criterion has a local minimum at either 18 or 19. A simulated example in which the underlying function $\sin(2\pi t/18.5)$ has period 18.5 is illustrated in Figure 9.

We can also evaluate a CV criterion at all the midpoints between two integers when the sample size is not too small. For a period of the form $q + 1/2$, stacking data in the usual way results in $2q$ distinct time points in each cycle, but, in comparison to the case of period q, only half as many observations at each time point. We propose

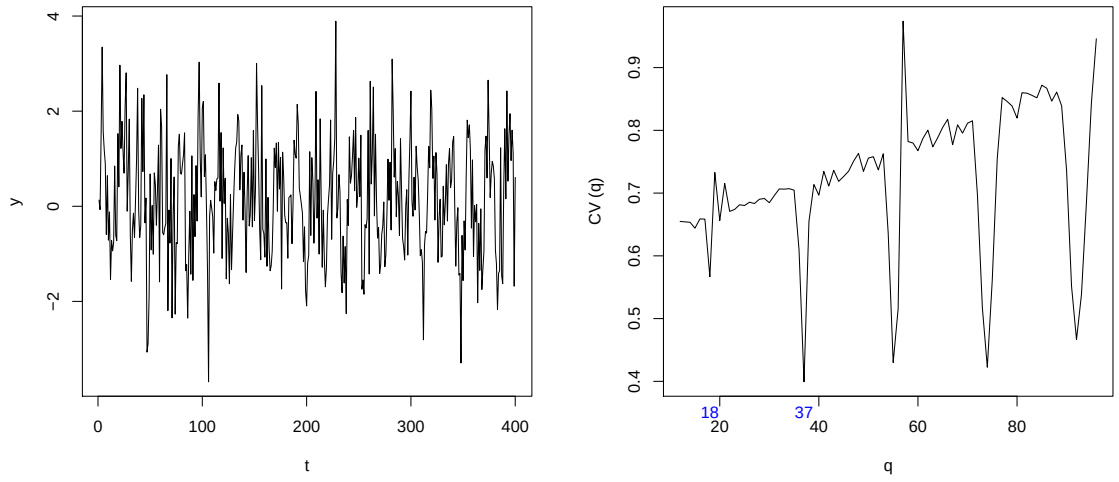

Figure 9. Left panel: time series plot with the period of the underlying function equal to 18.5 and $n = 400$. Right panel: the corresponding CV criterion plot.

in such a case to leave out a cycle, compute sample means of the remaining data, and then smooth these sample means by a three-point moving average. Then the data in the deleted cycle may be predicted by these smoothed means. This method was applied to the same data as in Figure 9. The modified CV criterion is shown in Figure 10, and now the CV curve is minimized at $q = 18.5$.

2.6.2. Concluding remarks

We have proposed a cross-validation period estimator for equally spaced data that are the sum of a periodic sequence and noise. The method is computationally simple and implicitly penalizes multiples of the smallest period. Given a particular period, or cycle length, a leave-out-one-cycle version of CV is used to compute an average squared prediction error. The cycle length minimizing this average squared error

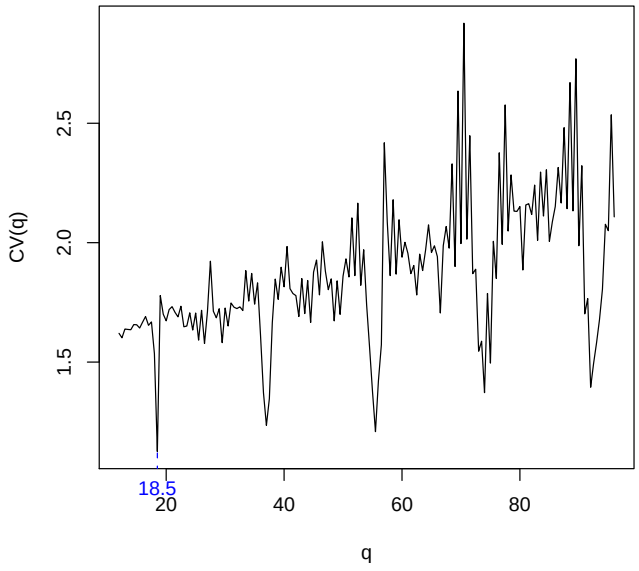

Figure 10. The CV criterion plot for evaluating non-integer period with the period of the underlying function equal to 18.5 and $n = 400$.

is the period estimator. Moreover, models corresponding to different periods may be ranked from best to worst by considering values of the objective function, thus, extending the possibilities of interpretation.

It is shown both theoretically and by simulation that the CV method has a much higher probability of choosing the correct model than it does in familiar cases where the considered models are nested. Our theory shows that the CV period estimator $\hat{p}$ is virtually consistent for large p, in that its asymptotic probability of equaling p increases monotonically to 1 as p becomes large. When $p = 15$ this probability is approximately 0.99. It is worth noting that the CV method has the advantage that it can easily deal with missing data, as long as the missing data are at random.

CHAPTER III

NONPARAMETRIC ESTIMATION OF MULTIPLE PERIODIC SEQUENCES

3.1. Introduction

A number of nonparametric methods for period estimation in univariate time series have been proposed recently, including Hall, Reimann and Rice (2000), Hall and Yin (2003), Hall and Li (2006), Genton and Hall (2007) and Hall (2008). In these papers, to estimate the period of a periodic function, an appropriate random spacing of time points is needed in order to ensure the consistency, in other words, unevenly spaced observations are needed. Considering that equally spaced data are prevalent in time series analysis, Sun, Hart and Genton (2011) proposed a cross-validation based nonparametric method for evaluating the periodicity when time points are evenly spaced and a periodic sequence is observed. In many real world problems, however, multivariate time series are available. For instance, several different variables are simultaneously recorded from a system under study, such as atmospheric temperature, pressure and humidity in meteorology; heart rate, blood pressure, respiration in physiology. A multivariate time series could also be recorded from one variable but in spatially extended systems, such as in studies of meteorology, where temperature recordings are obtained from probes or satellites at different spatial locations. In this chapter, we present a method of estimating the periods of multiple periodic sequences when data are observed at evenly spaced time points. The basic idea is to borrow information from other correlated sequences to improve the period estimation of interest.

Let d be the number of the sequences. We consider the following model:

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, n, \tag{3.1}$$

where $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{d,t})^{\mathrm{T}} = (X_t, \mathbf{Z}_t)^{\mathrm{T}}$ is the observed vector with $\mathbf{Z}_t = (Y_{2,t}, \ldots, Y_{d,t})$, $\boldsymbol{\mu}_t = (\mu_{1,t}, \ldots, \mu_{d,t})^{\mathrm{T}}$ is an unknown constant vector and the error vectors $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{d,t})^{\mathrm{T}}$ for $t = 1, \ldots, n$ are independent and identically distributed with mean zero, and covariance matrix Σ which can be partitioned into

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^{\mathrm{T}} & \Sigma_{22} \end{pmatrix}.$$

It is assumed that $\mu_{s,i} = \mu_{s,i+mp_s}$ for $i = 1, \ldots, p_s$, $m = 1, 2, \ldots$, $s = 1, \ldots, d$, and integers $p_1, \ldots, p_d$, each of which is at least 2. Of interest is estimating p_1, the period of the first sequence $X_1, \ldots, X_n$.

One motivation for our methodology is the study of periodicity of El Niño effects, which are defined as sustained increases of at least 0.5 °C in average sea surface temperatures over the east-central tropical Pacific Ocean. Sun, Hart and Genton (2011) have applied their method to a series of sea surface temperatures to estimate the El Niño period. In addition to sea surface temperatures, another series, sea level pressures were also measured at the same time points. We will investigate whether the El Niño period estimation would be improved by borrowing information from the pressure series.

Our method of estimating the period of interest is also based on cross-validation (CV). The CV estimator proposed by Sun, Hart and Genton (2011) for one periodic sequence evaluates a candidate period q by first "stacking," at the same time point, all data which are separated in time by a multiple of q, and then computing a "leave-out-one-cycle" version of the variance for each of the q stacks of data. In other words,

if we have k complete cycles where $k = n/q$, the stacked means of the other $k-1$ cycles are used to predict the ones left out and then the averaged prediction errors are computed. For multiple periodic sequences, suppose we are interested in estimating p_1 in model (3.1). Instead of the stacked means, we propose to use the conditional means, i.e. conditional on other correlated sequences, to predict the left out cycle in cross-validation. The prediction error will tend to be smallest when the period candidate q_1 equals p_1. Since cross-validation is typically used as a model selection tool, we will also consider model selection criteria for multiple periodic sequences other than CV in estimating a period.

Sun, Hart and Genton (2011) have described the asymptotic behavior of the one sequence CV method. They showed that when p is sufficiently large, the period estimator $\hat{p}_1$ is virtually consistent, in the sense that $\lim_{n\to\infty} P(\hat{p}_1 = p_1)$ increases to 1 as p_1 increases. For the multivariate CV, we show that it has the same asymptotic properties, but with simulations we show that for finite samples, stronger correlation between sequences, better period estimation of other correlated sequences and use of more correlated sequences all lead to substantial improvements in estimating the period of interest.

The rest of this chapter is organized as follows. Section 3.2 describes the cross-validation method of estimating the period and sequence values for multiple periodic sequences, while Section 3.3 discusses asymptotic properties of the method. Simulations motivated by real data applications are reported in Section 3.4. Concluding remarks are provided in Section 3.6 and a derivation of MLEs for bivariate sequence means is in Appendix B.

3.2. Methodology

Suppose we observe multiple sequences at equally spaced time points from the model (3.1), where for each $s = 1, \ldots, d$, the sequence $\mu_{s,1}, \ldots, \mu_{s,n}$ is periodic with (smallest) period p_s and the $\boldsymbol{\varepsilon}_t$s are independent and identically distributed random vectors with zero means and covariance matrix Σ. We propose a methodology for estimating p_1, the period of interest, in Section 3.2.1 and an alternative based on Akaike's information criterion (AIC) in Section 3.2.2. Besides period estimation, we also discuss the estimation of sequence values for multiple periodic sequences.

3.2.1. Multivariate cross-validation method for period estimation

For one periodic sequence, Sun, Hart and Genton (2011) proposed the cross-validation (CV) method for period estimation. Let q be a candidate integer period with $2 \leq q \leq M_n$, where M_n is of smaller order than $\sqrt{n}$. For each $i = 1, \ldots, q$, they constructed an estimator of μ_i by stacking all data that are separated in time by a multiple of q. So, at time points i the data are $X_i, X_{i+q}, \ldots, X_{i+qk_{q,i}}$, where $k_{q,i}$ is the largest integer such that $i + qk_{q,i} \leq n$. For each relevant q, i and j, define $X_{qij} = X_{i+(j-1)q}$. Let

$$\mathrm{CV}(q) = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} (X_{qij} - \bar{X}_{qi}^{j})^2, \tag{3.2}$$

where $\bar{X}_{qi}^{j}$ is the average of $X_{qi\ell}$, $\ell = 1, \ldots, k_{q,i}$, excluding X_{qij}. A period estimator $\hat{p}$ is defined to be the minimizer of $\mathrm{CV}(q)$ for $2 \leq q \leq M_n$.

Now, suppose we observe multiple sequences at equally spaced time points from the model (3.1). Let q_s be a candidate integer period of the s-th sequence. For the first sequence, let $q = q_1$ for simplicity and then $\bar{X}_{qi}^{j}$ is the average of $X_{qi\ell}$ excluding X_{qij}, $i = 1, \ldots, q$, $\ell = 1, \ldots, k_{qi}$ and $i + qk_{q,i} \leq n$. Define the averages for the sequences for $s = 2, \ldots, d$ in a similar way. For $i = 1, \ldots, q_s$, let $\bar{Y}_{s,q_s i}$ be the average

of $Y_{s,q_s i\ell}$, $\ell = 1, \ldots, k_{q_s,i}$ and $i + q_s k_{q_s,i} \leq n$. Then define the residual for each of the sequences to be $e_{s,q_s i\ell} = Y_{s,q_s i\ell} - \bar{Y}_{s,q_s i}$. Now, by defining the residual separately for each sequence, we have a residual vector $\mathbf{e}_t$ at each time point t, $t = 1, \ldots, n$. Let

$$\mathrm{CV}_d(q) = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} \left[X_{qij} - \left\{ \bar{X}_{qi}^{j} + \Sigma_{12} \Sigma_{22}^{-1} \mathbf{e}_{qij} \right\} \right]^2, \tag{3.3}$$

where $\mathbf{e}_{qij} = \mathbf{e}_{i+(j-1)q}$. We define a period estimator $\hat{p}_1$ to be the minimizer of $\mathrm{CV}_d(q)$.

The case of two sequences is a special case of the CV criterion in (3.3). If we observe two sequences $\{X_t : t = 1, \ldots, n\}$ and $\{Z_t : t = 1, \ldots, n\}$, the criterion (3.3) can be simplified as

$$CV_2(q) = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} \left[X_{qij} - \left\{ \bar{X}_{qi}^{j} + \frac{\sigma_1}{\sigma_2} \rho \, e_{qij} \right\} \right]^2, \tag{3.4}$$

where $e_{qij} = e_{i+(j-1)q}$, $i = 1, \ldots, q$, $j = 1, \ldots, k_{q,i}$, $r = 1, \ldots, q_2$ and $\ell = 1, \ldots, k_{q_2,r}$. Then, we propose to estimate p_1 in the following way:

S1. Apply the CV method (3.2) for one periodic sequence in Sun, Hart and Genton (2011) to each sequence and get the period estimates $\hat{p}^{(0)}$ and $\hat{p}_2$.

S2. Estimate the sequence values of each periodic sequence at the estimated periods. Given that the period is q, Sun, Hart and Genton (2011) showed that the maximum likelihood estimates of the sequence values $\mu_1, \ldots, \mu_q$ are the means $\bar{X}_{q1}, \ldots, \bar{X}_{qq}$, as defined above, when the errors are assumed to be i.i.d. Gaussian.

S3. Subtract the estimated sequence values from the observations, compute the sample covariance matrix from residual vectors, and obtain $\hat{\sigma}_1$, $\hat{\sigma}_2$ and $\hat{\rho}$.

S4. Construct $\bar{X}_{qi}^{j} + \frac{\hat{\sigma}_1}{\hat{\sigma}_2} \hat{\rho} \; e_{qij}$ to predict X_{qij} and compute the averaged squared prediction errors in (3.4).

S5. Choose q to minimize (3.4) and obtain the period estimate $\hat{p}^{(1)}$ of the first sequence.

S6. Repeat Step 2-5 and choose the period estimate $\hat{p}_1 = \hat{p}^{(2)}$.

When $\boldsymbol{\varepsilon}_t$ has an elliptical distribution, the best linear predictor of X_t given Z_t is $\mu_{1,t} + \frac{\sigma_1}{\sigma_2}\rho\varepsilon_{2,t}$. This motivates the predictor $\bar{X}^j_{qi} + \frac{\hat{\sigma}_1}{\hat{\sigma}_2}\hat{\rho}\, e_{qij}$ of X_{qij}. Note that when $\rho = 0$, equation (3.4) reduces to equation (3.2) for the one sequence case and the second sequence is not used.

3.2.2. Model selection criteria for multiple periods estimation

Cross-validation is typically used as a model selection tool. Sun, Hart and Genton (2011) discussed period estimation for one sequence from a model selection point of view, since in fact each candidate period q corresponds to a model for the sequence consisting of the q parameters $\mu_1, \ldots, \mu_q$. Similarly, for multiple periodic sequences, if we assume that the error vectors in model (3.1) are i.i.d. multivariate normal, then Akaike's information criterion has the form

$$\mathrm{AIC}(q_1, \ldots, q_d) = -2\log L(\boldsymbol{\mu}_t, \Sigma | \mathbf{Y}_t, t = 1, \ldots, n) + 2\sum_{s=1}^{d} q_s. \tag{3.5}$$

Minimizing $\mathrm{AIC}(q_1, \ldots, q_d)$ provides estimators of $p_1, \ldots, p_d$, but we need to plug in the MLEs of the sequence means. For the one sequence case, it is straightforward to verify that the MLEs of $\mu_1, \ldots, \mu_q$ are the stacked means $\bar{Y}_{q1}, \ldots, \bar{Y}_{qq}$. For multiple correlated sequences, instead of simple stacked means for each sequence, the MLEs depend on the covariance matrix Σ. The proof for the bivariate case is in Appendix B. We can see that only when $\rho = 0$ are the MLEs the usual stacked means. For $\rho \neq 0$, however, the MLEs depend on the value of ρ and the observations from both of the sequences.

3.3. Asymptotic Properties of the Bivariate CV Method

Sun, Hart and Genton (2011) showed the virtual consistency of the CV estimator, that is, $\lim_{n\to\infty} P(\hat{p} = p)$ increases to 1 as p increases. To describe the asymptotic behavior of the one sequence CV period estimator, they have shown that $P(\hat{p} = p)$ is asymptotically equal to

$$P\left(\bigcap_{1\le m\le M_n/p} \{\mathrm{CV}_1(mp) - \mathrm{CV}_1(p)\} > 0\right),$$

and then proved that $P(\mathrm{CV}_1(mp) - \mathrm{CV}_1(p) \le 0)$ is exponentially small when p is sufficiently large. For the bivariate case, suppose we observe bivariate sequences from the following model:

$$\begin{cases} X_t = \mu_{1,t} + \varepsilon_{1,t}, \\ Z_t = \mu_{2,t} + \varepsilon_{2,t}, \end{cases}$$

where $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}_t \sim$ i.i.d. $N(\mathbf{0}, \Sigma)$, and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

Let $\tilde{X}_q^{(t)}$ be the leave-out-one-cycle predictor of X_t using only data from the first sequence. The prediction error in the one sequence case is

$$X_t - \tilde{X}_q^{(t)} = \varepsilon_{1,t} + (\mu_{1,t} - \tilde{X}_q^{(t)}).$$

In the two sequences case, by using the true error term $\varepsilon_{2,t}$, the ideal prediction error is

$$X_t - \left(\tilde{X}_q^{(t)} + \frac{\sigma_1}{\sigma_2}\rho\varepsilon_{2,t}\right) = \varepsilon_{1,t} - \frac{\sigma_1}{\sigma_2}\rho\varepsilon_{2,t} + (\mu_{1,t} - \tilde{X}_q^{(t)}).$$

For $\delta_t = \varepsilon_{1,t} - \frac{\sigma_1}{\sigma_2}\rho\varepsilon_{2,t}$,

$$\mathrm{Var}(\delta_t) = \sigma_1^2 + \frac{\sigma_1^2}{\sigma_2^2}\rho^2\sigma_2^2 - 2\frac{\sigma_1}{\sigma_2}\rho\sigma_1\sigma_2\rho = \sigma_1^2(1-\rho^2) < \mathrm{Var}(\varepsilon_{1,t}).$$

Therefore, by using the second sequence, we can better predict the error term.

Now,

$$\begin{aligned}\mathrm{CV}_2(q) &= \frac{1}{n}\sum_{t=1}^{n}\left(X_t - \tilde{X}_q^{(t)} - \frac{\sigma_1}{\sigma_2}\rho\varepsilon_{2,t}\right)^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}(X_t - \tilde{X}_q^{(t)})^2 - \frac{2}{n}\sum_{t=1}^{n}(X_t - \tilde{X}_q^{(t)})\frac{\sigma_1}{\sigma_2}\rho\varepsilon_{2,t} + \frac{1}{n}\sum_{t=1}^{n}\frac{\sigma_1^2}{\sigma_2^2}\rho^2\varepsilon_{2,t}^2 \\ &= \mathrm{CV}_1(q) - \frac{2}{n}\frac{\sigma_1}{\sigma_2}\rho\sum_{t=1}^{n}(X_t - \tilde{X}_{1q}^{(t)})\varepsilon_{2,t} + \frac{1}{n}\frac{\sigma_1^2}{\sigma_2^2}\rho^2\sum_{t=1}^{n}\varepsilon_{2,t}^2.\end{aligned}$$

Then,

$$\mathrm{CV}_2(mp) - \mathrm{CV}_2(p) = \mathrm{CV}_1(mp) - \mathrm{CV}_1(p) + \frac{2}{n}\frac{\sigma_1}{\sigma_2}\rho\sum_{t=1}^{n}\varepsilon_{2,t}(\tilde{X}_{mp}^{(t)} - \tilde{X}_p^{(t)}).$$

Let $A_n = \frac{2}{n}\frac{\sigma_1}{\sigma_2}\rho\sum_{t=1}^{n}\varepsilon_{2,t}(\tilde{X}_{mp}^{(t)} - \tilde{X}_p^{(t)}) = \frac{2}{n}\frac{\sigma_1}{\sigma_2}\rho\sum_{t=1}^{n}\varepsilon_{2,t}(\tilde{\varepsilon}_{1mp}^{(t)} - \tilde{\varepsilon}_{1p}^{(t)})$. Now, $\varepsilon_{2,t}$ is independent of $\tilde{\varepsilon}_{1mp}^{(t)} - \tilde{\varepsilon}_{1p}^{(t)}$, both are normally distributed and $\mathrm{Var}(\tilde{\varepsilon}_{1mp}^{(t)} - \tilde{\varepsilon}_{1p}^{(t)}) = O(\frac{1}{n})$. Then, proving that the extra term A_n is negligible can be done in the same way $A_{q,n}$ was treated in the proof of Theorem 1 in Appendix A.

Therefore the asymptotic distribution of the CV estimator does not change by adding another correlated sequence. For finite samples, however, it does improve the estimation which will be shown by simulation in Section 3.4.

3.4. Simulation

3.4.1. Bivariate case

To study whether another correlated periodic sequence improves the period estimation, we performed simulations with an intermediate value of the period, i.e., $p_1 = 43$, which is the estimate of the El Niño period, and $p_2 = 59$ which is our estimate of the sea level pressure period. Let $\sigma_1 = \sigma_2 = 1$ in model (3.1). Since the sea surface temperatures are negatively correlated with the sea level pressures, we consider three

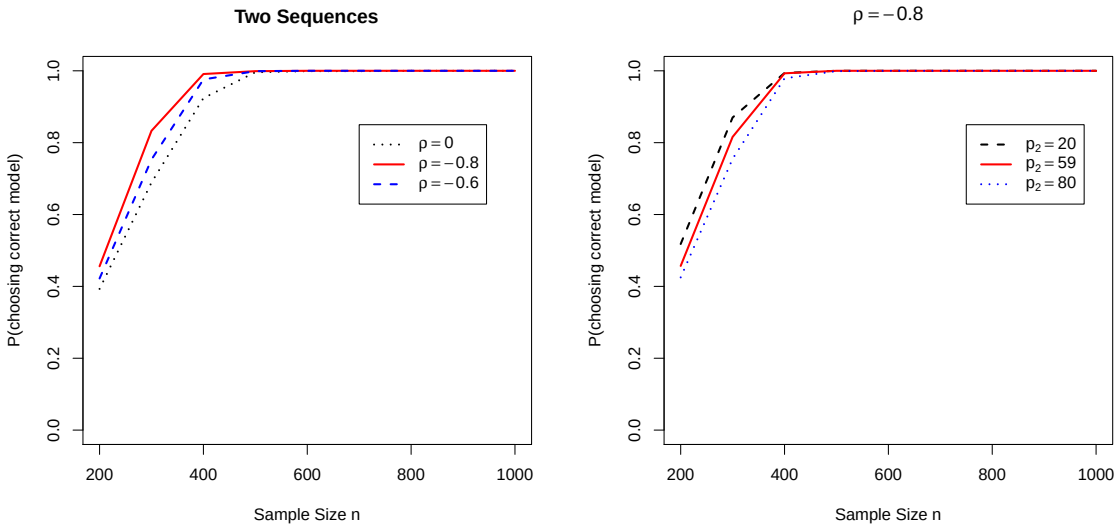

Figure 11. Left panel: $P(\hat{p}_1 = p_1)$ for the bivariate CV method at different correlation levels. Right panel: $P(\hat{p}_1 = p_1)$ for the bivariate CV method at different values of p_2.

cases, $\rho_1 = 0, -0.6, -0.8$. For all cases, the set over which the objective function $\mathrm{CV}_2(q)$ was searched was taken to be $\{12, 13, \ldots, 96\}$, and use the algorithm S1-S5 proposed in Section 3.2.1 to estimate p_1. The number of replications of each setting is 1000.

The left panel of Figure 11 shows how the probabilities of choosing $p_1 = 43$ increase as the sample size n increases for each $\rho = 0, -0.6, -0.8$ when $p_2 = 59$. It is clear that the convergence is faster for larger correlation. We can also see that for finite sample, the second sequence does improve the estimation of p_1 and more improvements could be obtained with stronger correlation. Then we consider different values of p_2 when $\rho = -0.8$. The probabilities of choosing $p_1 = 43$ are shown in the right panel of Figure 11 for $p_2 = 20, 59, 80$. For a fixed sample size n, the probability of choosing $p_1 = 43$ is higher for smaller value of p_2 due to more available cycles of

the second sequence. Thus, for a fixed value of ρ, how much the second series helps for estimating p_1 depends on how well p_2 is estimated as well.

Now, suppose $p_1 = p_2 = 43$, $\rho = -0.8$ and $\sigma_1 = 1$. We consider different error levels of the second sequence. Figure 12 shows the probabilities of choosing $p_1 = 43$ for the different error levels $\sigma_2 = 0.5, 1, 1.5$. The convergence is slower for larger error levels, which is further evidence that how well p_2 is estimated plays a role in the estimation of p_1.

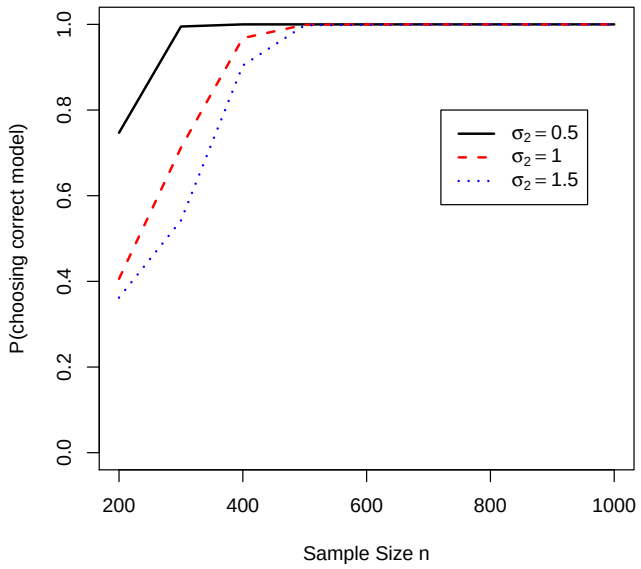

Figure 12. $P(\hat{p}_1 = p_1)$ for the bivariate CV method at different error levels of the second sequence. The values of σ_1 and ρ are 1 and -0.8, respectively, in each case.

3.4.2. Trivariate case

Suppose we observe three periodic sequences from model (3.1). Let $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\rho_{12} = -0.8$ which is the same as the correlation in the bivariate simulation, $\rho_{13} = 0.6$ and $\rho_{23} = -0.5$. Initially we let $p_1 = 43$, $p_2 = p_3 = 59$. The probabilities of choosing $p_1 = 43$ are shown in the left panel of Figure 13. With three sequences, the convergence is faster than either one or two sequences which shows that one more correlated sequence further improves the estimation of p_1, although the improvement from 1 to 2 sequences is apparently larger than improvements from 2 to 3, etc. We also consider a situation with $p_2 = p_3 = 20$. The right panel of Figure 13 shows that the convergence is faster for $p_2 = p_3 = 20$ than for $p_2 = p_3 = 59$. This is for the same reason as in the two sequences case, i.e., more cycles are available for a smaller period when n is fixed.

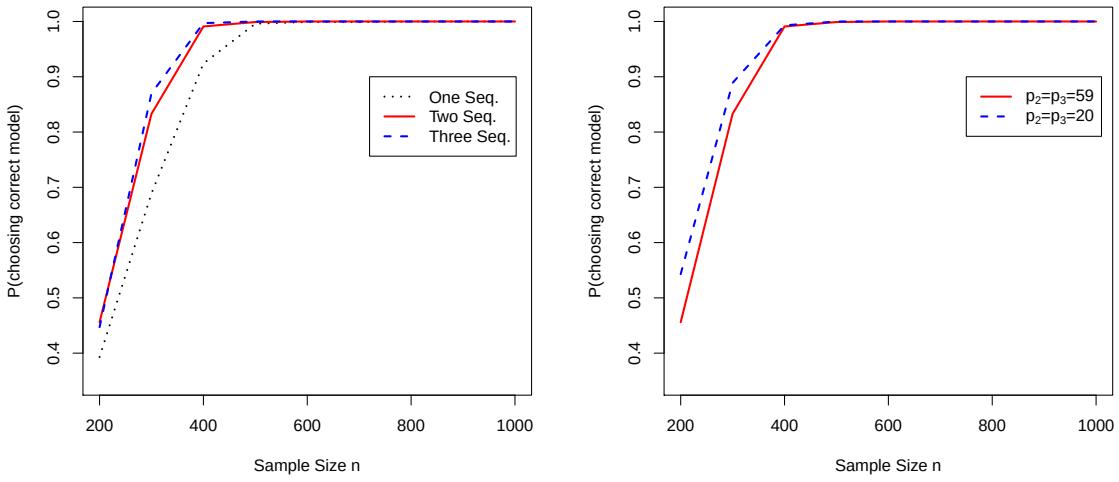

Figure 13. Left panel: $P(\hat{p}_1 = p_1)$ for the univariate, bivariate and trivariate CV methods. Right panel: $P(\hat{p}_1 = p_1)$ for the trivariate CV method when $p_2 = p_3 = 59$ and $p_2 = p_3 = 20$.

3.4.3. AIC for bivariate case

For the bivariate case, the AIC in (3.5) becomes

$$\mathrm{AIC}(q_1, q_2) = \frac{1}{1-\rho^2}\left\{\frac{1}{\sigma_1^2}\sum_{t=1}^{n}(X_t-\mu_{1,t})^2 + \frac{1}{\sigma_2^2}\sum_{t=1}^{n}(Z_t-\mu_{2,t})^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{t=1}^{n}(X_t-\mu_{1,t})(Z_t-\mu_{2,t})\right\} + 2(q_1+q_2),$$

where q_1 and q_2 are the period candidates for p_1 and p_2 and we choose the period estimates as the minimizer of $\mathrm{AIC}(q_1, q_2)$. Again, let $p_1 = 43$, $p_2 = 59$, $\sigma_1 = \sigma_2 = 1$ and $\rho = -0.8$. The set for q_1 and q_2 over which the $\mathrm{AIC}(q_1, q_2)$ was searched was taken

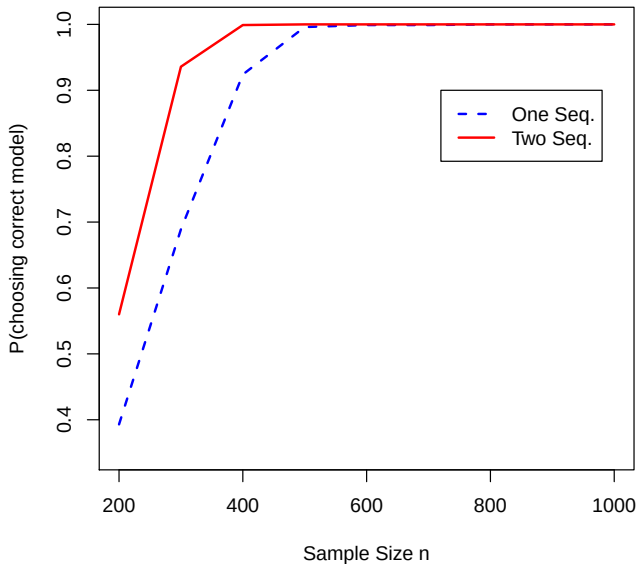

Figure 14. $P(\hat{p}_1 = p_1)$ for the bivariate case by the AIC method comparing to the CV method for one sequence.

to be $\{12, 13, \ldots, 96\}$, and the number of replications of each setting is 1000. In the simulation study, we assume the parameters σ_1, σ_2 and ρ are known, and the mean

parameters were estimated by the closed form in Appendix B, which is a function of the observations, ρ and the ratio of σ_1 and σ_2. Then the criterion AIC(q_1, q_2) was minimized over the set $\{12, 13, \ldots, 96\} \times \{12, 13, \ldots, 96\}$. The probabilities of choosing $p_1 = 43$ are shown in Figure 14. Similarly, it is clear that the convergence is faster for the two sequences case.

3.5. Applications

El Niño is a phenomenon associated with warmer than normal sea surface temperatures (SST), which can have profound effects on local climate. The data that has been

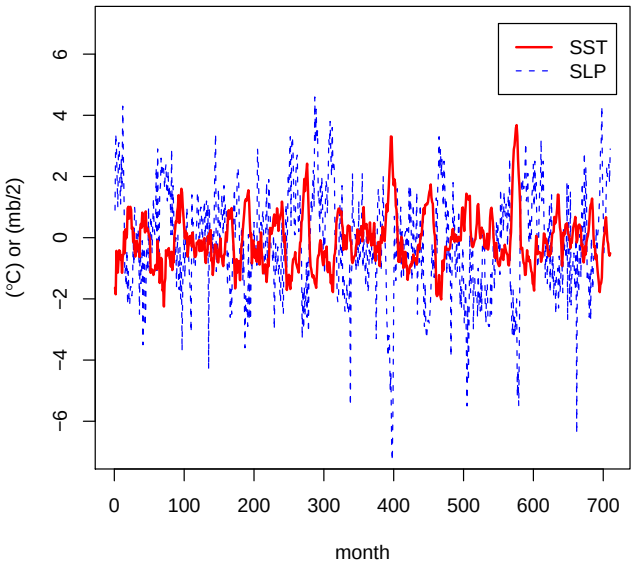

Figure 15. The time series plot of sea surface temperatures and sea level pressures from 1950 to 2009: the red solid line denotes the SST and the blue dashed line represents the SLP.

analyzed by Sun, Hart and Genton (2011) consist of the area-average SST anomalies

over the eastern equatorial Pacific, monthly from January 1950 to February 2009. In addition to SST, another series, monthly sea level pressures (SLP), was also measured at the same time points. As can been seen in Figure 15, the two sequences are negatively correlated.

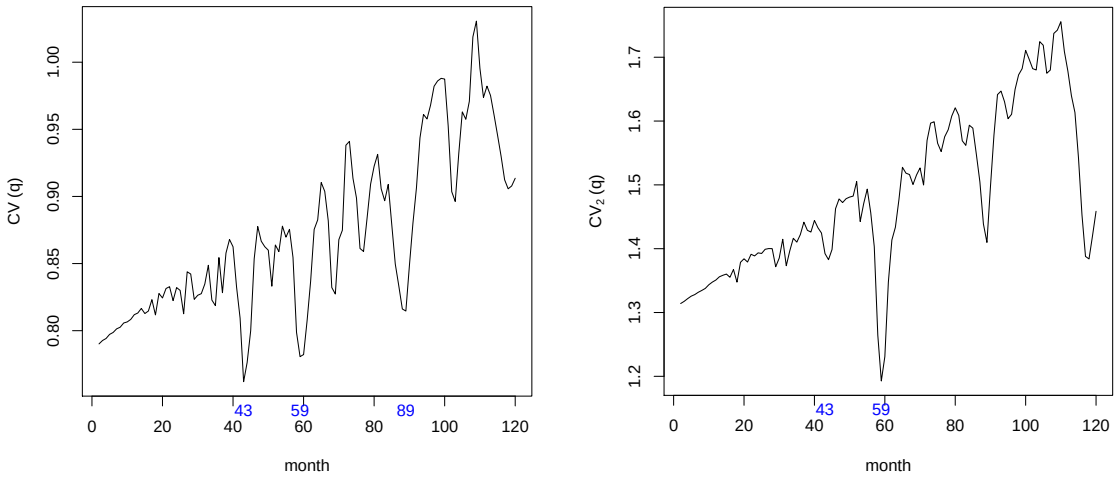

Figure 16. Left panel: the plot of the CV curve for estimating the period of the SST sequence. Right panel: the plot of the CV curve for estimating the period of the SST sequence when using both SST and SLP sequences.

The cross-validation method proposed by Sun, Hart and Genton (2011) yields a period estimate of $\hat{p}_1 = 43$ months; see the left panel of Figure 16 for the CV curve. The 43 estimated means plot is shown in Figure 17. Notice that the CV curve is also locally minimized at 59 months. Interestingly, by adding the SLP sequence to help estimate the period of the SST sequence, $\hat{p}_1$ becomes 59 months; see the right panel of Figure 16 for the CV curve, where the minimum switches from 43 months to 59 months. The 59 estimated means plot is shown in the right panel of Figure 17. Comparing with the 43 estimated means plot shown in the left panel of Figure 17, the

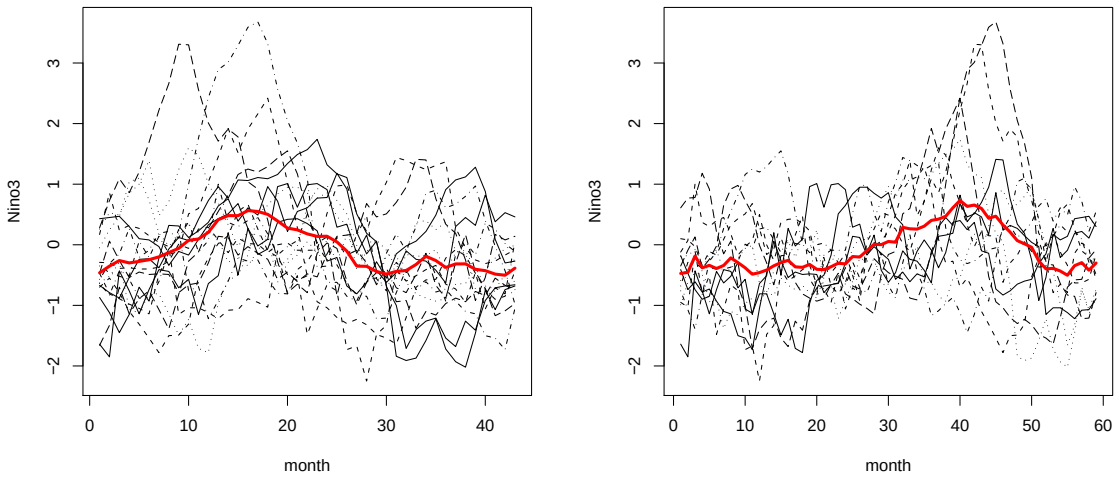

Figure 17. Left panel: a plot of the 16 1/2 cycles of data (corresponding to a period of 43 months) superimposed on each other. The 43 estimated means are connected by the red line. Right panel: a plot of the 12 cycles of data (corresponding to a period of 59 months) superimposed on each other. The 59 estimated means are connected by the red line.

peaks of each cycle seem to better coincide than when cycles are folded at the period 43 months. To measure the proportion of variability in a dataset that is accounted for by the statistical model, we define the coefficient of determination R^2 in the same way as the R-squared measure in linear regression,

$$R^2 = \frac{\sum_{i=1}^{n}(\bar{Y}_{\hat{p}i} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

where $\bar{Y}_{\hat{p}i}$ is the stacked mean at time point i and $i = 1, \ldots, n$. For the SST sequence, the model with $\hat{p}_1 = 59$ has a larger R-squared value, $R^2 = 0.166$, comparing to $R^2 = 0.144$ when $\hat{p}_1 = 43$. This provides another evidence that the model with $\hat{p}_1 = 59$ is a better one.

3.6. Discussion

In this chapter, we have proposed a cross-validation period estimator for multiple equally spaced periodic sequences. Sharing a similar idea with the CV method for one sequence, a leave-out-one-cycle version of CV is used to compute an average squared prediction error given a particular period, or cycle length. The multivariate CV method uses the conditional means, i.e. conditional on other correlated sequences, to predict the left out cycle in cross-validation. In this way, the period estimation for a sequence X has been improved by borrowing information from other sequences with which X is correlated. In theory, we show that the asymptotic behavior of the bivariate CV is the same as the CV for one sequence. In our simulation studies, however, it is shown that for finite samples, the better the periods of the other correlated sequences are estimated, the more substantial improvements can be obtained in estimating the period of interest. We have also shown that more correlated sequences lead to more improvements, although the improvement from 1 to 2 sequences is apparently larger than improvements from 2 to 3, etc. In addition to the CV method, we also considered a model selection criterion, AIC, to estimate the period for multiple periodic sequences. Similarly, for finite samples, a simulation study shows an improvement in period estimation from using information in a correlated sequence. The asymptotic properties of the AIC method will need further exploration.

CHAPTER IV

FUNCTIONAL BOXPLOTS*

4.1. Introduction

Functional data analysis is an attractive approach to study complex data in statistics. In many statistical experiments, the observations are functions by nature, such as temporal curves or spatial surfaces, where the basic unit of information is the entire observed function rather than a string of numbers. Such functional data appear in many fields, including meteorology, biology, medicine, and engineering. Human growth curves, weather station temperatures, gene expression signals, medical images, and human speech are all real life examples; see e.g. Dryden and Mardia (1998), Fletcher et al. (2004), and Ramsay and Silverman (2005).

To analyze functional data, researchers often used mathematical models, among which Ramsay and Silverman (2005) provided various parametric methods while Ferraty and Vieu (2006) developed detailed nonparametric techniques. Quantile regression, as a popular model-based method, has been widely used, and many economic applications were discussed by Fitzenberger et al. (2002). In contrast to model-based analysis, visualization methods often help to display the data, highlight their characteristics and reveal interesting features. For functional data, Hyndman and Shang (2010) proposed two graphical methods with outlier detection capability: the functional bagplot and the functional highest density region boxplot, both of which are

*Reprinted with permission from "Functional Boxplots" by Sun, Y. and Genton, M. G., 2011. Journal of Computational and Graphical Statistics, 20, 316-334, Copyright [2011] by American Statistical Association.

based on the first two robust principal component scores. They apply the bivariate bagplot (Rousseeuw et al., 1999) to the first two robust principal component scores, and then map the features of the bagplot into the functional space. In this chapter, we aim to develop visualization tools for functional data directly in the functional space rather than in the feature space that requires principal component analysis techniques.

It is well-known that the boxplot is a graphical method for displaying five descriptive statistics: the median, the first and third quartiles, and the non-outlying minimum and maximum observations. A boxplot may also indicate which observations, if any, can be considered as outliers. First introduced by Tukey (1970) and Tukey (1977, pp. 39-43) in exploratory data analysis, boxplots have evolved into a straightforward but informative method in data interpretation. The first step to construct a boxplot is the data ordering. In the univariate setting, the ranking is simply from the smallest observation to the largest. However, multivariate ordering is much more complicated and has attracted considerable interest over the years. To generalize order statistics or ranks to the multivariate setting, different versions of data depth have been introduced to measure how deep (central) or outlying an observation is. Examples of data depth include the Mahalanobis depth (Mahalanobis, 1936), the Tukey halfspace location depth (Tukey 1975), the Oja depth (Oja, 1983), the simplicial depth (Liu, 1990), the majority depth (Singh, 1991), and the likelihood depth (Fraiman and Meloche, 1999). Vardi and Zhang (2000) proposed an L_1-depth which can be extended to functional data. Febrero et al. (2007, 2008) have reviewed a series of functional depths, such as the functional depth of Fraiman and Muniz (2001), the functional depth of Cuevas et al. (2006) and the random projection functional depth of Cuevas et al. (2007).

For functional data, López-Pintado and Romo (2009) recently introduced a no-

tion of band depth (BD). It allows for ordering a sample of curves from the center outward and, thus, introduces a measure to define functional quantiles and the centrality or outlyingness of an observation. Having the ranks of curves, the functional boxplot is a natural extension of the classical boxplot and is an appealing visualization tool for functional data.

This chapter is organized as follows. Section 4.2 explains the definition of band depth for functional data and its modified version. Section 4.3 illustrates the construction of functional boxplots and enhanced functional boxplots, as well as the associated outlier detection rule. Simulation results on the performance of our outlier detection method are reported in Section 4.4. The visualization capabilities of the functional boxplots are demonstrated in Section 4.5 when applied to classical functional data and a space-time dataset. A discussion is provided in Section 4.6.

4.2. Band Depth for Functional Data

In functional data analysis, each observation is a real function $y_i(t)$, $i = 1, \ldots, n$, $t \in \mathcal{I}$, where $\mathcal{I}$ is an interval in $\mathbb{R}$. A band depth for functional data provides a method to order all the sample curves. Indeed, we can compute the band depths of all the sample curves and order them according to decreasing depth values. Let $y_{[i]}(t)$ denote the sample curve associated with the ith largest band depth value. We view $y_{[1]}(t), \ldots, y_{[n]}(t)$ as order statistics, with $y_{[1]}(t)$ being the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ being the most outlying curve. The implication is that a smaller rank is associated with a more central position with respect to the sample curves. The order statistics induced by a band depth start from the most central sample curve and move outwards in all directions. Therefore, they are different from the usual order statistics which are simply ordered from the

smallest sample value to the largest.

With this basic idea, López-Pintado and Romo (2009) introduced the band depth concept through a graph-based approach. The graph of a function $y(t)$ is the subset of the plane $G(y) = \{(t, y(t)) : t \in \mathcal{I}\}$. The band in $\mathbb{R}^2$ delimited by the curves $y_{i_1}, \ldots, y_{i_k}$ is $B(y_{i_1}, \ldots, y_{i_k}) = \{(t, x(t)) : t \in \mathcal{I}, \min_{r=1,\ldots,k} y_{i_r}(t) \leq x(t) \leq \max_{r=1,\ldots,k} y_{i_r}(t)\}$. Let J be the number of curves determining a band, where J is a fixed value with $2 \leq J \leq n$. If $Y_1(t), \ldots, Y_n(t)$ are independent copies of the stochastic process $Y(t)$ generating the observations $y_1(t), \ldots, y_n(t)$, the population version of the band depth for a given curve $y(t)$ with respect to the probability measure P is defined as

$$BD_J(y, P) = \sum_{j=2}^{J} BD^{(j)}(y, P) = \sum_{j=2}^{J} P\{G(y) \subset B(Y_1, \ldots, Y_j)\},$$

where $B(Y_1, \ldots, Y_j)$ is a band delimited by j random curves. The sample version of $BD^{(j)}(y, P)$ is obtained by computing the fraction of the bands determined by j different sample curves containing the whole graph of the curve $y(t)$. In other words, $BD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq n} I\{G(y) \subseteq B(y_{i_1}, \ldots, y_{i_j})\}$, where $I\{\cdot\}$ denotes the indicator function. The implication is that by computing the fraction of the bands containing the curve $y(t)$, the bigger the value of band depth, the more central position the curve has. Then, the sample band depth of a curve $y(t)$ is

$$BD_{n,J}(y) = \sum_{j=2}^{J} BD_n^{(j)}(y). \tag{4.1}$$

Instead of considering the indicator function, López-Pintado and Romo (2009) also proposed a more flexible definition, the modified band depth (MBD), by measuring the proportion of time that a curve $y(t)$ is in the band: $MBD_n^{(j)}(y) = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq n} \lambda_r\{A(y; y_{i_1}, \ldots, y_{i_j})\}$, where $A_j(y) \equiv A(y; y_{i_1}, \ldots, y_{i_j}) \equiv \{t \in$

$\mathcal{I} : \min_{r=i_1,\ldots,i_j} y_r(t) \leq y(t) \leq \max_{r=i_1,\ldots,i_j} y_r(t)\}$ and $\lambda_r(y) = \lambda(A_j(y))/\lambda(\mathcal{I})$, if λ is the Lebesgue measure on $\mathcal{I}$. If $y(t)$ is always inside the band, the modified band depth degenerates to the band depth in (4.1).

Because the modified band depth takes the proportion of times that a curve is in the band into account, it avoids having too many depth ties and is more convenient to obtain the most representative curves in terms of magnitude. The band depth is more dependent on the shape of curves often yielding ties, thus it can be used to obtain the most representative curves in terms of shape. Consequently, there are two types of outliers: magnitude outliers and shape outliers. In general, magnitude outliers are distant from the mean and shape outliers have a pattern different from the other curves.

A sample median function is a curve from the sample with largest depth value, defined by $\arg\max_{y\in\{y_1,\ldots,y_n\}} BD_{n,J}(y)$. If there are ties, the median will be the average of the curves maximizing depth.

Although the number of curves determining a band, j, could be any integer between 2 and J, the order of curves induced by band depth is very stable in J. To avoid computational issues, we use $J = 2$, and for simplicity, we write $BD_n^{(2)}$ as BD and $MBD_n^{(2)}$ as MBD in the sequel.

Figure 18 provides a simple example with $n = 4$ curves on how to compute BD and MBD in practice. When $J = 2$, there are 6 possible bands delimited by 2 curves. For instance, the grey area in Figure 18 is the band delimited by $y_1(t)$ and $y_3(t)$. We can see that the curve $y_2(t)$ completely belongs to the band, but $y_4(t)$ only partly does. We define that a curve is contained in a band even if this curve is on the border of the band. Then $BD(y_2) = 5/6 = 0.83$ since only the band delimited by $y_3(t)$ and $y_4(t)$ does not completely contain the curve $y_2(t)$ and $BD(y_4) = 3/6 = 0.5$ as it is only completely contained in the bands delimited by itself and another curve. Similarly,

we could compute $BD(y_1) = 0.5$ and $BD(y_3) = 0.5$. To compute MBD, note that the curve $y_2(t)$ is always contained in the five bands, hence $MBD(y_2) = 0.83$, the same value as BD. In contrast, the curve $y_4(t)$ only belongs to the band in grey 40% of the time, thus $MBD(y_4) = (3 + 0.4 + 0.4)/6 = 0.63$ by definition. For the other two curves, $MBD(y_1) = 0.5$ and $MBD(y_3) = 0.7$.

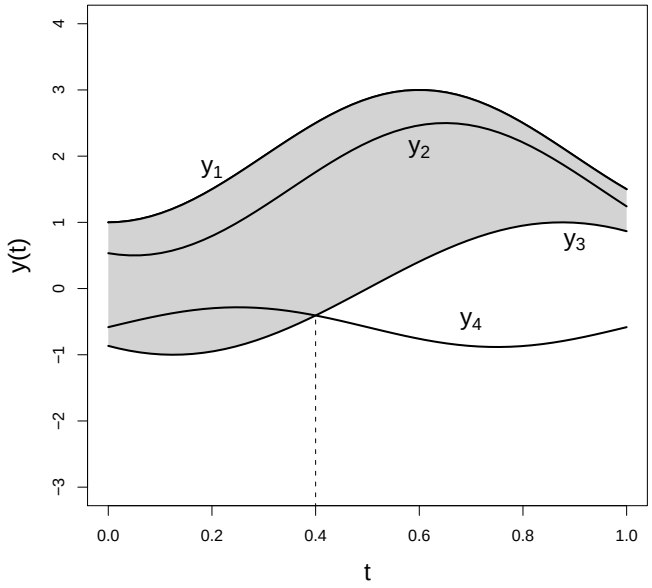

Figure 18. An example of BD and MBD computation: the grey area is the band delimited by $y_1(t)$ and $y_3(t)$. The curve $y_2(t)$ completely belongs to the band, but $y_4(t)$ only partly does.

4.3. Construction of Functional Boxplots

In the classical boxplot, the box itself represents the middle 50% of the data. An interesting idea that can be extended to functional data is the concept of central region introduced by Liu et al. (1999). The band delimited by the α proportion $(0 < \alpha < 1)$ of deepest curves from the sample is used to estimate the α central

region. In particular, the sample 50% central region is

$$C_{0.5} = \{(t, y(t)) : \min_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t) \leq y(t) \leq \max_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t)\},$$

where $\lceil n/2 \rceil$ is the smallest integer not less than $n/2$. The border of the 50% central region is defined as the envelope representing the box in a classical boxplot. Thus, this 50% central region is the analog to the "inter-quartile range" (IQR) and gives a useful indication of the spread of the central 50% of the curves. This is a robust range for interpretation because the 50% central region is not affected by outliers or extreme values, and gives a less biased visualization of the curves' spread. There is also a curve in the box that indicates the median $y_{[1]}(t)$, or the most central curve which has largest band depth value. The median curve is also a robust statistic to measure centrality.

The "whiskers" of the boxplot are the vertical lines of the plot extending from the box and indicating the maximum envelope of the dataset except the outliers. Thus, we need to identify the outliers first. Again, we extend the 1.5 times IQR empirical outlier criterion to the functional boxplot. The fences are obtained by inflating the envelope of the 50% central region by 1.5 times the range of the 50% central region. Any curves outside the fences are flagged as potential outliers. It is worth noting that when each curve is simply a point, the functional boxplot degenerates to a classical boxplot. We suggest the constant factor 1.5 as in a classical boxplot, but we leave to the user the possibility of modifying it.

Now that the pieces of the functional boxplot have been identified, we illustrate its construction on a dataset used by Hyndman and Shang (2010) to demonstrate their functional bagplot. The data consist of monthly sea surface temperatures (SST) measured in degrees Celsius over the east-central tropical Pacific Ocean and are shown in Figure 19. In this case, each curve represents one year of observed SST in degrees Cel-

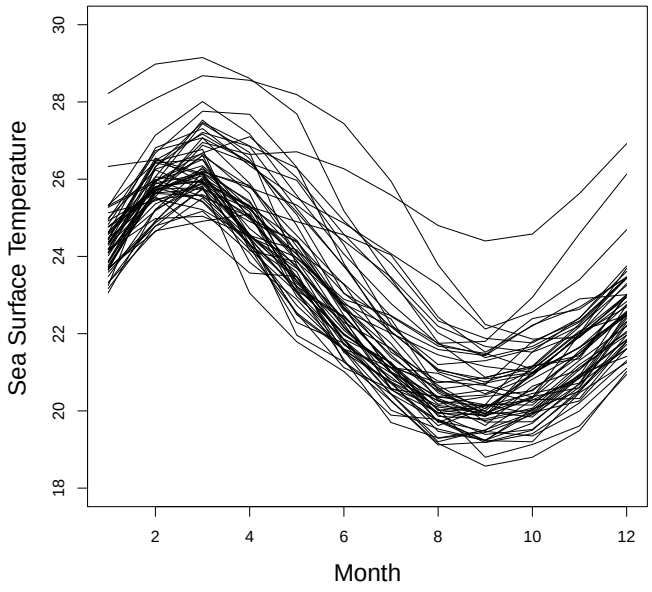

Figure 19. Data of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean from 1951 to 2007.

sius from January 1951 to December 2007. In our functional boxplot (Figure 20 (a)), only the median curve and the flagged outliers are real observations. The border of the box in the middle denotes the envelope of the 50% central region and the minimum and maximum provide the range of non-outlying envelope. To show this difference, we use blue curves to denote envelopes, a black curve to represent the median curve and red dashed curves to indicate outlier candidates. Thus, instead of having five summary statistics as in a classical boxplot, the functional boxplot has the envelope of the central 50% region, the median curve and the maximum non-outlying envelope as descriptive statistics.

As can be seen from Figure 20 (a) and Figure 20 (c), the two methods display the same median curve in this example, but slightly different outlier detection results.

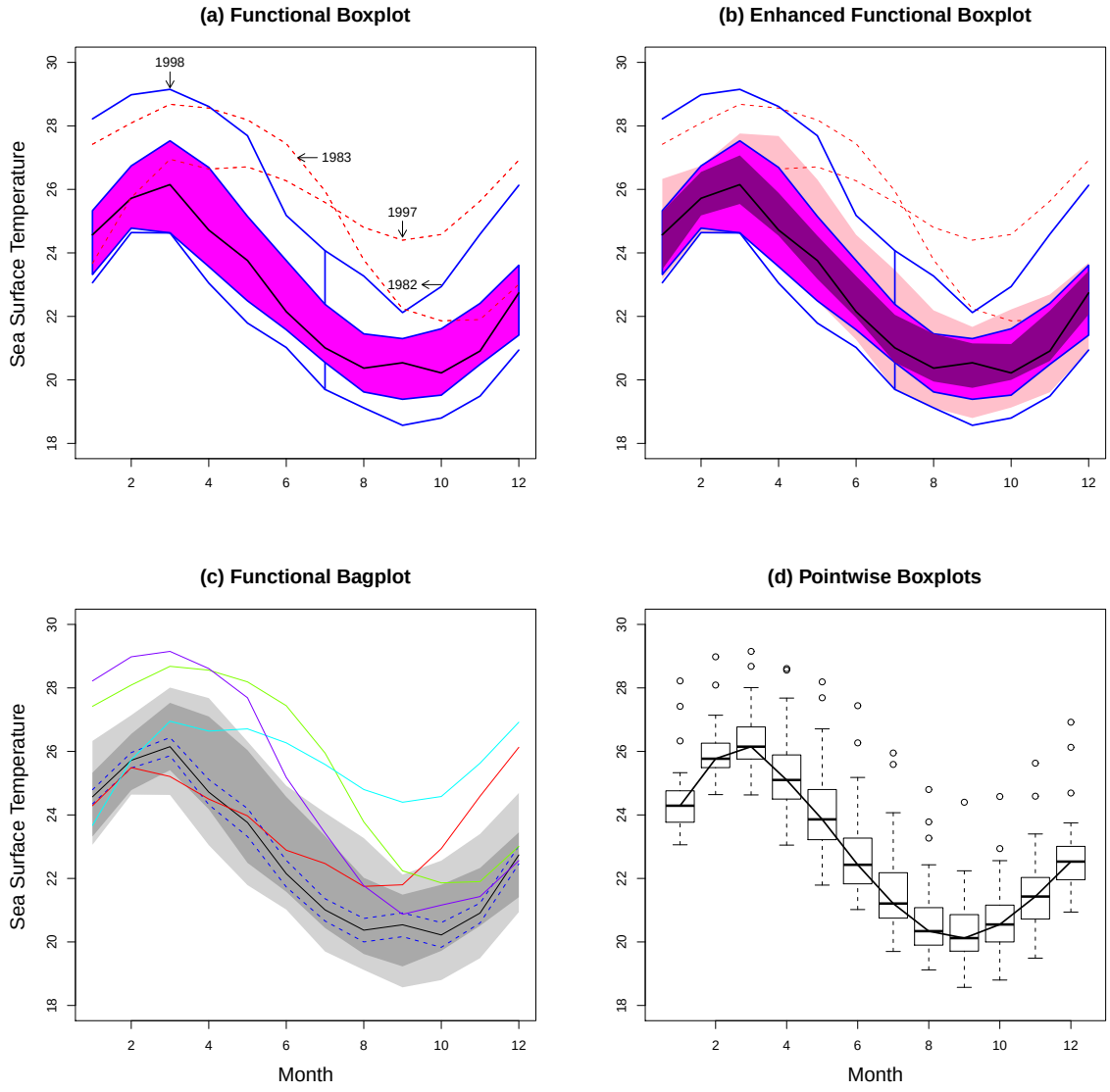

Figure 20. (a): the functional boxplot of SST with blue curves denoting envelopes, and a black curve representing the median curve. The red dashed curves are the outlier candidates detected by the 1.5 times the 50% central region rule. (b): the enhanced functional boxplot of SST with dark magenta denoting the 25% central region, magenta representing the 50% central region and pink indicating the 75% central region. (c): the functional bagplot of SST. (d): the pointwise boxplots of SST with medians connected by a black line.

Our functional boxplot detects two outliers by using MBD: the years 1983 and 1997. In addition, the year 1982 from September to December and the year 1998 from January to June are viewed as being part of the maximum envelope. The information discovered by the functional boxplot that September 1982 to December 1983 and January 1997 to June 1998 are abnormal is in close agreement with the recent major El Niño events reported by Dioses et al. (2002). Similarly, the functional bagplot of Hyndman and Shang (2010) detects the years 1982-1983 and 1997-1998 as outliers. For functional data, such as these sea surface temperatures, there will necessarily be dependence in time. This is why the outliers come in adjacent years. Considering that the dependence in time may affect outlier detection performance, we allow the constant factor 1.5 to be adjustable in practice.

By introducing the concept of central regions, the functional boxplot can be generalized to an enhanced functional boxplot shown in Figure 20 (b). Besides the 50% central region, the 25% and 75% central regions are provided as well. We have implemented a function `fbplot` in `R` (R Development Core Team, 2010) to produce functional boxplots and enhanced functional boxplots. It is available as supplemental material on the *JCGS* website.

One may think of using the most intuitive approach, the pointwise boxplots shown in Figure 20 (d), which do not treat each curve as one observation. Obviously, such an approach has lost the information of the curves' shapes. In general, the central regions provided by pointwise boxplots are narrower than those given by the functional boxplot, thus many more points would be detected as outliers. By comparing these two types of boxplots, we see that the functional median could be equivalent to the medians in pointwise boxplots only if all the points on the functional median curve are the pointwise 50% quantiles simultaneously. This is rarely true for functional data, especially when curves are very irregular. Specifically, in the above

sea surface temperatures example, outliers are detected for each month without taking the annual trend into account. One may connect those monthly outliers from the same year, but it is very difficult to visualize the whole outlying yearly curve and there are cases where only one or two monthly observations within one year are relatively extreme. Furthermore, using the connected pointwise medians (the middle black line in Figure 20 (d)) as the most representative curve is not very sensible since it smoothes out too many monthly features of a typical yearly temperature curve and is no longer a true curve of the sample.

It is important to note that the box, the whiskers and the median can reveal useful information about a functional dataset by looking at their position, size, length, and even the shape of the box or the median curve. Moreover, the spacings between the different parts of the box help indicate the degree of skewness in the data and identify outliers.

4.4. Simulation Studies

Hyndman and Shang (2010) proposed the functional bagplot and the functional highest density region (HDR) boxplot of which both can detect outliers. The former obtains the outer region (the "fence") by inflating the inner region (the "bag") by a constant factor 2.58 and the latter needs to prespecify the coverage probability of the outlying region. We will focus on comparing our functional boxplot with their functional bagplot since the empirical outlier rule we have proposed obtains the outer region (the "fence") by inflating the inner region (the "envelope") by 1.5 times the range of the 50% central region. We prefer not to have to prespecify the coverage probability of the outlying region in case there is no outlier or the fraction of outliers is unknown.

To further compare our functional boxplot with the principal component (PC) based functional bagplot and assess their performance for outlier detection, we have generated curves from different models introducing either magnitude outliers or shape outliers. The model structures are similar to those in López-Pintado and Romo (2009), but with different parameter values. Some of these models were already considered in Fraiman and Muniz (2001).

Model 1 is a basic one without contamination shown in the left panel of Figure 21. Model 2, model 3 and model 4 have magnitude outliers while model 5 has shape contamination as shown in the left panel of Figure 22. Model details are described as follows:

1. Model 1 is $X_i(t) = g(t) + e_i(t), 1 \leq i \leq n$, with mean $g(t) = 4t$, $t \in [0, 1]$ and where $e_i(t)$ is a stochastic Gaussian process with zero mean and covariance function $\gamma(s, t) = \exp\{-|t - s|\}$.
2. Model 2 includes a symmetric contamination: $Y_i(t) = X_i(t) + c_i\sigma_i K$, where c_i is 1 with probability q and 0 with probability $1 - q$, K is a contamination size constant and σ_i is a sequence of random variables independent of c_i taking values 1 and -1 with probability $1/2$.
3. Model 3 is partially contaminated: $Y_i(t) = X_i(t) + c_i\sigma_i K$, if $t \geq T_i$ and $Y_i(t) = X_i(t)$, if $t < T_i$, where T_i is a random number generated from a uniform distribution on $[0, 1]$.
4. Model 4 is contaminated by peaks: $Y_i(t) = X_i(t) + c_i\sigma_i K$, if $T_i \leq t \leq T_i + \ell$, and $Y_i(t) = X_i(t)$ otherwise, where T_i is a random number from a uniform distribution in $[0, 1 - \ell]$.
5. Model 5 considers shape contamination with different parameters in the covariance function $\gamma(s, t) = k \exp\{-c|t-s|^{\mu}\}$. The basic model 1, $X_i(t) = g(t)+e_{1i}(t)$

has parameter values $k = 1$, $c = 1$ and $\mu = 1$ for the covariance function of $e_{1i}(t)$. To generate irregular curves, let $Y_i(t) = g(t) + e_{2i}(t)$, where $e_{2i}(t)$ is a Gaussian process with zero mean and covariance function parameters $k = 8$, $c = 1$ and $\mu = 0.2$. The contaminated model is given by $Z_i(t) = (1 - c_i)X_i(t) + c_i Y_i(t)$, $1 \leq i \leq n$, where c_i is 1 with probability q and 0 with probability $1 - q$.

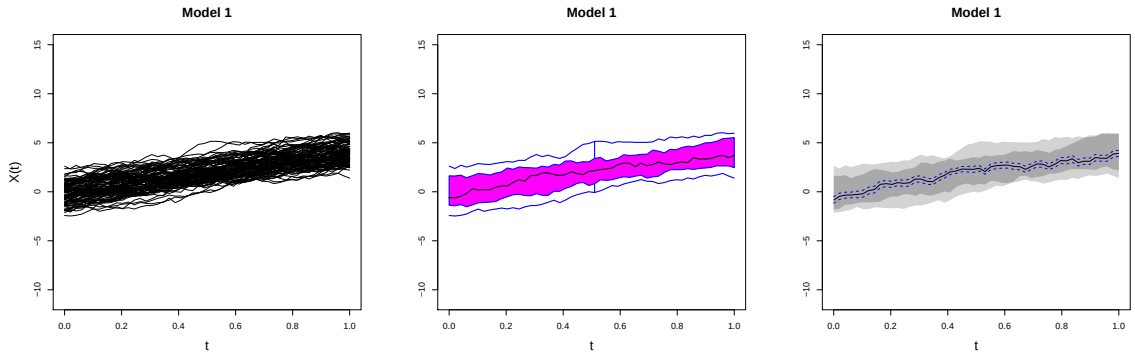

Figure 21. Left panel: curves generated from model 1. Middle panel: the corresponding functional boxplot. Right panel: the corresponding functional bagplot.

In the simulation studies, we generate $n = 100$ curves with parameters $q = 0.1$, $K = 8$, $\ell = 3/49$, and compute depth values by MBD, the more flexible version of band depth. Figure 21 and Figure 22 show the difference of outlier detection between our band depth based functional boxplots and the functional bagplots of Hyndman and Shang (2010) based on the first two PC scores. For this particular generated dataset, both methods work equally well on the first three models and the first two PCs of the robust covariance matrices explain 87.0%, 85.0% and 89.3% of the total variation, respectively. However, the PC based functional bagplot only detects one outlier in model 4, and in model 5 it misses most of the outliers and falsely detects one non-outlying curve. For these two models, the first two PCs explain only 78.3%

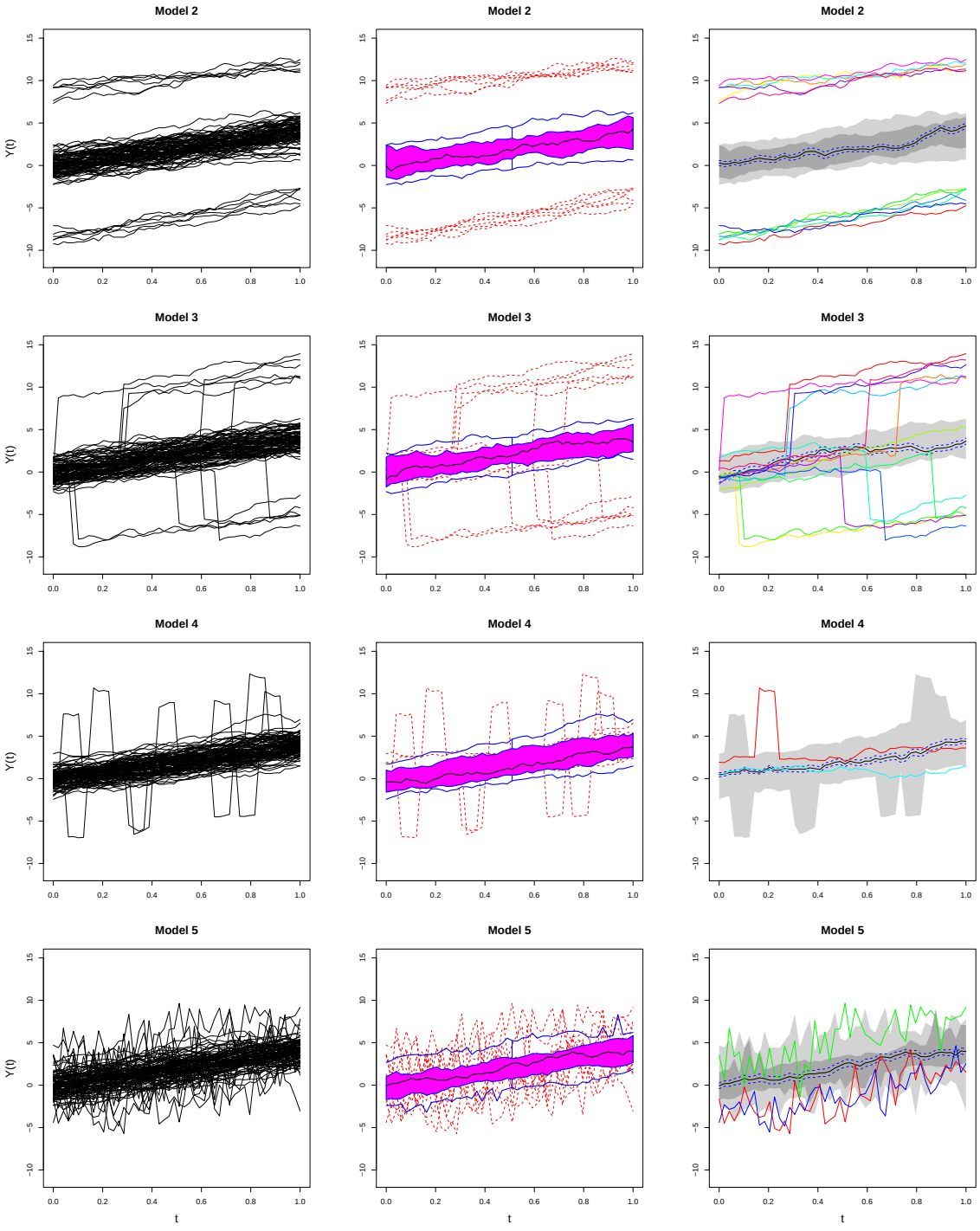

Figure 22. Left panels: curves generated from each contaminated model. Middle panels: the corresponding functional boxplots. Right panels: the corresponding functional bagplots.

and 77.5% of the total variation, respectively, which are smaller than those in models 1 to 3. Thus, using only the first two PCs is sometimes a potential drawback of the functional bagplot.

To assess the variability of the outlier detection methods, we are interested in the distribution of two quantities: p_c, the percentage of correctly detected outliers (number of correctly detected outliers divided by the total number of outlying curves), and p_f, the percentage of falsely detected outliers (number of falsely detected outliers divided by the total number of non-outlying curves).

For model 1, the basic model without outliers, we estimate the percentage, p_0, that each of the two methods detects no outliers, and obtain the distribution of the percentage $\hat{p}_f$ with 1000 replications and 100 curves. The percentage, the mean and standard deviation of $\hat{p}_f$ are shown in Table 3. For models 2 to 5, we obtain the distribution of the two percentages $\hat{p}_c$ and $\hat{p}_f$ with 1000 replications and 100 curves. The means and standard deviations are shown in Table 4. A good performance is defined as high correct detection percentages p_0 and p_c, but a low false detection percentage p_f. As can be seen, overall the functional boxplot method works better than the functional bagplot except for model 3, where, however, the two methods are not significantly different considering the variation. Focusing on the models 1, 4 and 5, the better performance of the functional boxplot method is obvious and significant. The simulation results show that the functional bagplot method is more likely to either miss a true outlier or falsely detect a non-outlying curve because it only depends on the first two principal components.

Notice that the peaks only appear during short intervals in model 4. By definition, BD would give small depth values for this type of outlying curves but MBD may not. Thus, an alternative would be to compute depth values by BD and to break the possible ties by their MBD values. In this way, simulation results show that the

mean of $\hat{p}_c$ could be increased to 95%.

Table 3: The percentage $\hat{p}_0$, the mean and standard deviation of the percentage $\hat{p}_f$ for the functional boxplot and functional bagplot with 1000 replications, 100 curves for model 1.

Method	$\hat{p}_0$	Mean($\hat{p}_f$)	SD($\hat{p}_f$)
Functional Boxplot	93.2	0.07	0.27
Functional Bagplot	24.4	2.42	6.24

Table 4: The mean and standard deviation (in the parentheses) of the percentage $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplots and functional bagplots with 1000 replications, 100 curves for models 2 to 5.

$\hat{p}_c$	Model 2	Model 3	Model 4	Model 5
Functional Boxplot	99.1(3.1)	83.7(13.9)	55.0(18.4)	78.6(15.3)
Functional Bagplot	99.5(5.5)	88.4(10.8)	18.6(15.7)	32.7(17.0)
$\hat{p}_f$	Model 2	Model 3	Model 4	Model 5
Functional Boxplot	0.03(0.19)	0.03(0.20)	0.05(0.27)	0.03(0.18)
Functional Bagplot	1.81(5.80)	1.51(4.59)	1.82(5.51)	1.66(5.13)

As another simulation study with harmonic signals, we simulated $n = 100$ curves of the form $Y_i(x) = (1 - c_i)\{a_{1i}\sin(t) + a_{1i}\cos(t)\} + c_i\{b_{1i}\sin(t) + b_{2i}\cos(t)\}$, where $0 < t < 2\pi$, c_i is 1 with probability 0.1 and 0 with probability 0.9. The coefficients a_{1i} and a_{2i} follow independent uniform distributions on [0,0.05], and b_{1i} and b_{2i} also follow independent uniform distributions but on [0.1,0.15]. This model (model 6) is similar to the third example studied in Hyndman and Shang (2010), but we introduce

outliers randomly with probability 0.1.

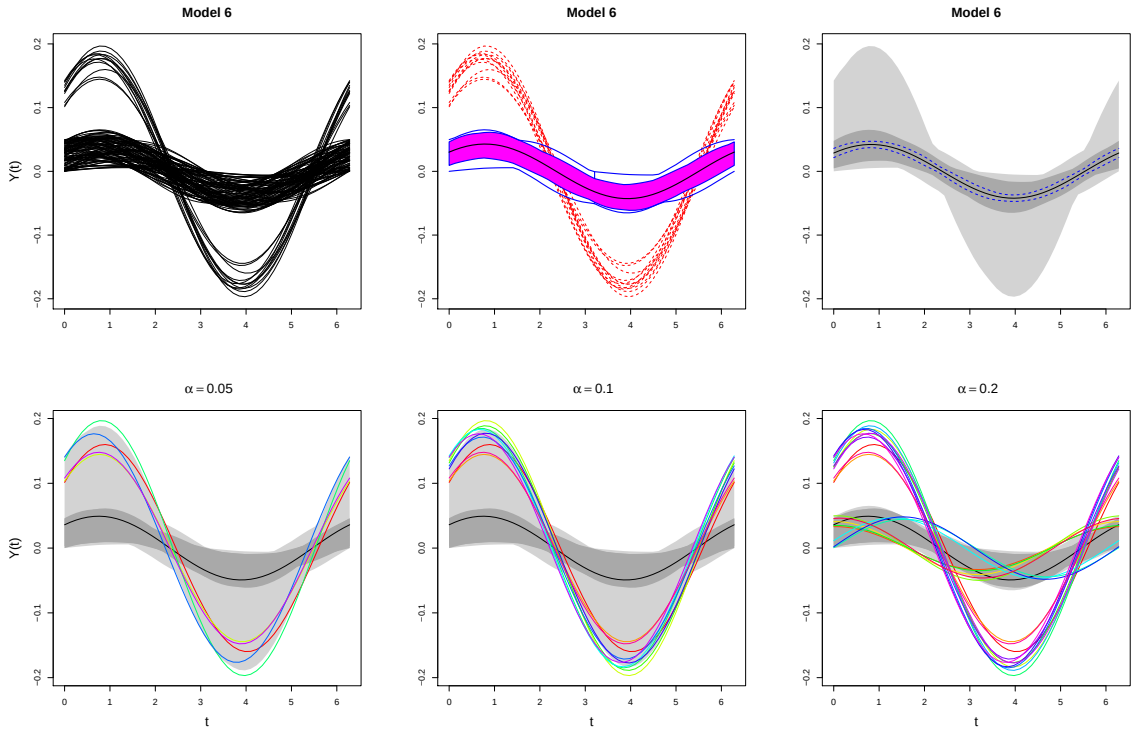

Figure 23. Top panels: the original curves generated from model 6, the corresponding functional boxplot and the functional bagplot. Bottom panels: the corresponding functional HDR boxplots for $\alpha = 0.05, 0.1, 0.2$, respectively.

For one particular generated dataset, the original curves and the corresponding functional boxplot and functional bagplot are shown in the top panels of Figure 23. Since the functional highest density region (HDR) boxplot needs to prespecify α, the coverage probability of the outlying region, the corresponding HDR boxplots for $\alpha = 0.05, 0.1, 0.2$ are shown in the bottom panels of Figure 23. For this dataset, the functional boxplot correctly detects all the outliers, but the functional bagplot fails to detect any. The three HDR boxplots clearly show that the outlier detection performance highly depends on the prespecified α. When α increases, more outliers

are detected but non-outlying curves are also more likely to be flagged as potential outliers at the same time.

Similarly, we obtain the distribution of the two percentage $\hat{p}_c$ and $\hat{p}_f$ for model 6 with 1000 replications and 100 curves. The means and standard deviations are reported in Table 5. The simulation results show that the functional boxplot also works better than the functional bagplot for model 6 and also better than the functional HDR boxplot even with the correctly prespecified outlier probability. For the functional HDR boxplots, the mean of $\hat{p}_c$ and $\hat{p}_f$ both increases as α increases. Hence, the outlier detection performance depends on the choice of α.

Table 5: The mean and standard deviation (in the parentheses) of the percentage $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplot, the functional bagplot and the HDR boxplots with 1000 replications, 100 curves for model 6.

Method	Functional Boxplot	Functional Bagplot	Functional HDR Boxplots		
			$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
$\hat{p}_c$	100(0.2)	72.8(42.4)	54.7(17.7)	90.7(12.9)	100(0.6)
$\hat{p}_f$	0(0)	0.48(4.50)	0.07(0.32)	1.41(1.75)	11.1(3.0)

Any outlier detection method should take care of both magnitude and shape outliers. However, to detect shape outliers not far from the median curve with lower density is not an easy task. The functional boxplot would be a good outlier detection method when outliers are either far away from the median in magnitude (models 2 to 4), or outlying in terms of shape but with some outlyingness in magnitude as well (models 5 and 6). However, it may miss outliers which are completely outlying in shape without showing any feature of magnitude outliers. This is where a density approach such as a functional highest density region boxplot can be useful, albeit the fact that the percentage of potential outliers must be known and the first two

PC scores must explain most of the variation. To illustrate this situation, we let the parameters a_{1i} and a_{2i} in model 6 follow independent uniform distributions on [0,0.1], and b_{1i} and b_{2i} follow independent uniform distributions on [0.1,0.12]. In this model (model 7), the parameters have the same values as the third example in Hyndman and Shang (2010), which make the outliers not very outlying in magnitude. The only difference is that we still simulate 100 curves and introduce outliers randomly with a probability of 0.1.

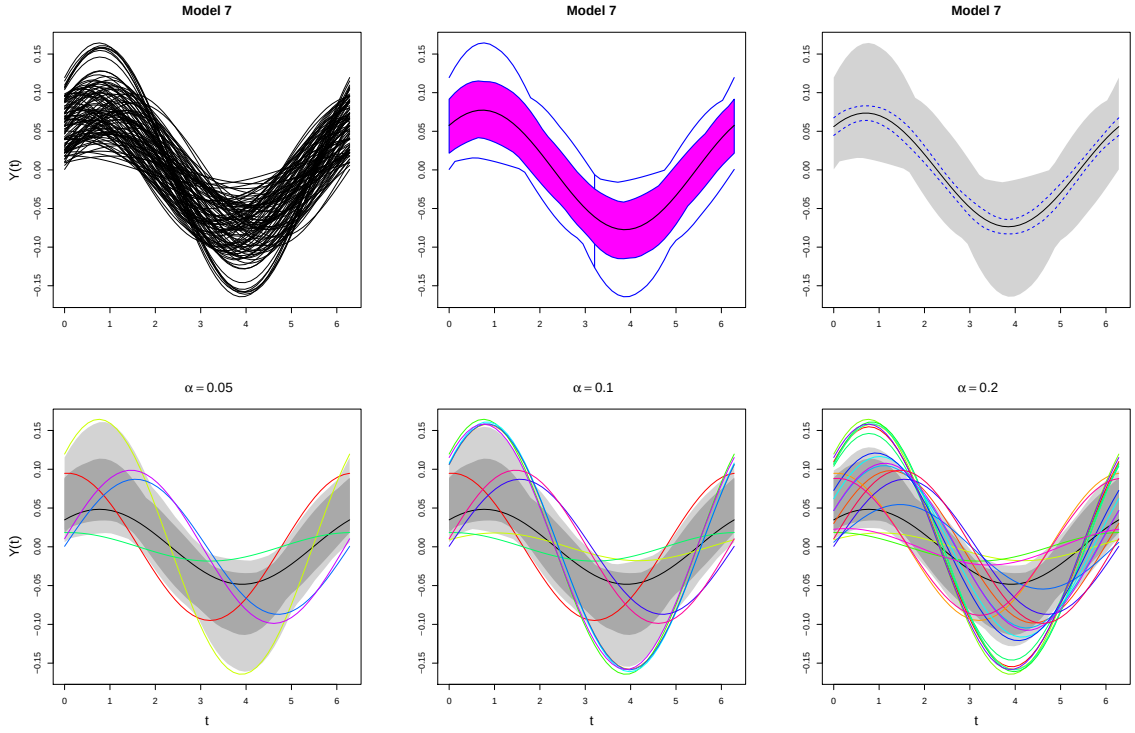

Figure 24. Top panels: the original curves generated from model 7, the corresponding functional boxplot and the functional bagplot. Bottom panels: the corresponding functional HDR boxplots for $\alpha = 0.05, 0.1, 0.2$, respectively.

For one particular generated dataset, the original curves and the corresponding functional boxplot and functional bagplot are shown in the top panels of Figure 24,

and the corresponding HDR boxplots for $\alpha = 0.05, 0.1, 0.2$ are shown in the bottom panels of Figure 24. For this dataset, the functional boxplot and functional bagplot fail to detect any of the outliers because the outlying curves are not sufficiently distant from the median. All three HDR boxplots detect some of the outliers but also flag other curves as potential outliers. As in model 6, when α increases, more and more outliers are detected. With 1000 replications and 100 curves, we obtain the distribution of the two percentage $\hat{p}_c$ and $\hat{p}_f$ for model 7. The means and standard deviations are reported in Table 6. It is shown that the functional boxplot fails to detect the outliers that are not far from the median, and the functional bagplot also fails most of the time. In contrast, the HDR boxplots can identify more such outliers but the correct detection rate is not high. For instance, the mean of $\hat{p}_c$ is only 17.5% even with the correctly prespecified $\alpha = 0.1$. A larger α could increase the correct detection rate, however, the false detection rate increases as well.

Table 6: The mean and standard deviation (in the parentheses) of the percentage $\hat{p}_c$ and $\hat{p}_f$ for the functional boxplot, the functional bagplot and the HDR boxplots with 1000 replications, 100 curves for model 7.

Method	Functional Boxplot	Functional Bagplot	Functional HDR Boxplots		
			$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
$\hat{p}_c$	0(0)	1.73(12.31)	8.25(19.39)	17.5(29.9)	33.0(38.5)
$\hat{p}_f$	0(0)	0.66(5.26)	5.07(1.15)	9.95(2.08)	19.7(3.3)

4.5. Applications

4.5.1. Children growth data

A strong point of the functional boxplot is its ability to display differences between populations without making any assumptions on the underlying statistical distribution. We start by applying the functional boxplot to the children growth data of Ramsay and Silverman (2005). The heights of 54 girls and 39 boys were measured at 31 unequally spaced ages from 1 year to 18 years. Within each population, the growth curves are monotonic and similar to a shifted version of each other. Thus we use the MBD because it is more suitable for magnitude outliers as we have discussed in Section 4.2.

Comparing the original curves to the functional boxplots in Figure 25, we see that the latter are very informative to compare the boys and girls data. The four blue curves and the black curve are the analog to the five summary statistics in a classical boxplot as we explained in the previous section. The median curves can be interpreted as the most representative observed patterns of children growth with age. In the functional boxplot for girls, we notice that there is one detected outlier candidate (red dashed curve), and girls reach lower height values at the end of the growth curves. Also, the shape of the boxes and the median curves depict that boys grew faster than girls between age 13 and 15 years. This information is difficult to obtain by simply looking at the original curves. In addition, one girl is detected as an outlier candidate whose growth curve is always higher in magnitude than the rest. In terms of the shape, this girl grew a little faster at her early age and stopped growing earlier but then still ended up taller than the others.

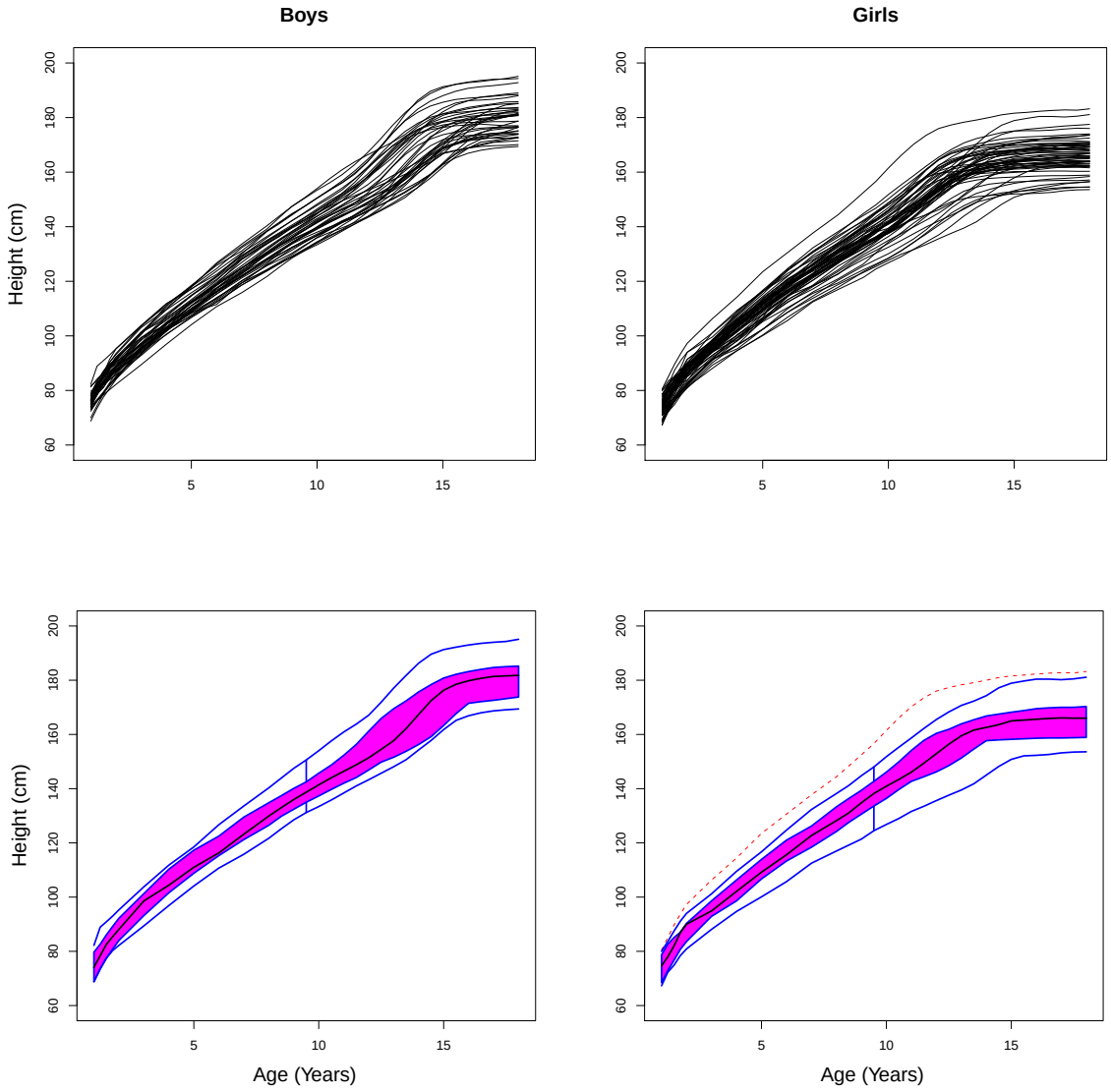

Figure 25. Top panels: the heights of 39 boys and 54 girls at 31 unequally spaced ages. Bottom panels: the corresponding functional boxplots of the children growth data using MBD.

4.5.2. Spatio-temporal precipitation data

Another feature of the functional boxplot is its ability to summarize information from complex data, such as space-time datasets. To illustrate this aspect, we use the observed annual total precipitation data for the coterminous U.S. from 1895 to 1997, provided by the Institute for Mathematics Applied to Geosciences at the web page (*http://www.image.ucar.edu/Data/US.monthly.met/*). There are 11,918 stations reporting precipitation at some time in this period. The observations are functional data since we have one time series with $p = 103$ yearly precipitation observations, or one curve, at each spatial location. Before we apply the functional boxplot to this complex dataset, we first need to perform smoothing to estimate each mean precipitation curve because the records of precipitation at each weather station are so variable. The original data were smoothed by a spline smoothing approach in a nonparametric regression model $y_j = f(x_j) + \varepsilon_j$, where ε_j i.i.d. $\sim N(0, \sigma^2)$, $j = 1, \ldots, p$. Spline smoothing uses all unique data points $x_1, \ldots, x_p$ as knots in the formulation of the cubic spline. Then the cubic spline estimator is obtained by minimizing $\sum_{j=1}^{p}\{y_j - f(x_j)\}^2 + \lambda_i \int \ddot{f}(x)^2\, dx$, where $\ddot{f}(x)$ is the second derivative of $f(x)$ and λ_i is the smoothing parameter of the ith curve. The smoothing parameters were estimated from the data by generalized cross-validation. Using functional boxplots to summarize and compare the annual precipitation for different climatic regions is an interesting application. Nine climatic regions for precipitation in the U.S. are defined by the National Climatic Data Center (NCDC) and shown in Figure 26. The number of stations is large for each region: the minimum number is 823 for the East North Central region and the maximum number is 2,084 for the South region. Blue dots denote stations with normal precipitation and red plus signs present potential outlying stations with respect to their respective climatic region detected by enhanced

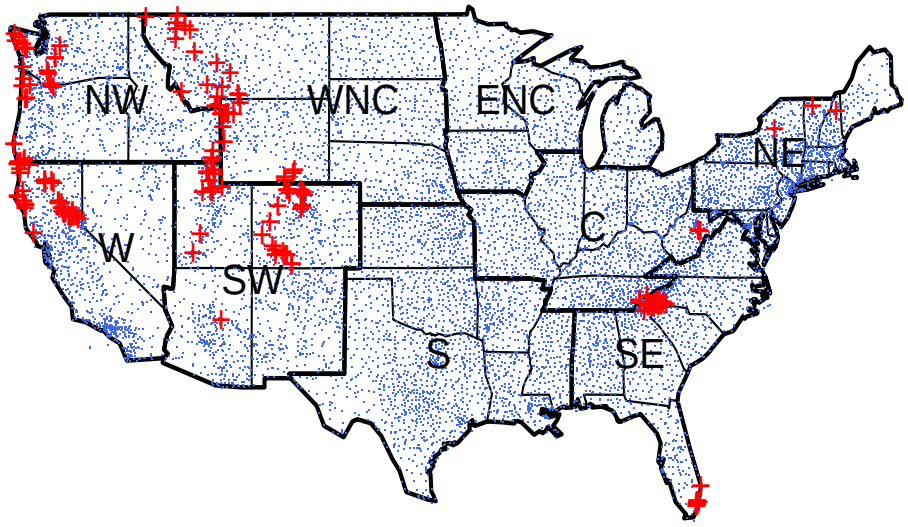

Figure 26. U.S. climatic regions for precipitation from NCDC with abbreviations for North East, East North Central, Central, South East, West North Central, South, South West, North West, and West. Blue dots denote stations with normal precipitation and red plus signs present potential outlying stations with respect to their respective climatic region detected by enhanced functional boxplots.

functional boxplots.

The nine enhanced functional boxplots based on MBD in Figure 27 reveal information about the different annual precipitation characteristics for different climatic regions. For each region, the global spatial outliers denoted by red dashed curves correspond to the red plus signs on the U.S. map in Figure 26.

There are mainly four areas of potential outliers within the U.S. shown in Figure 26. Two of them are located along the Rocky mountains in the West and the Appalachian mountains in the East with different patterns from the other locations in the corresponding climatic regions. In addition, certain amounts of potential outliers appear along the west coast with higher precipitation which can be clearly seen

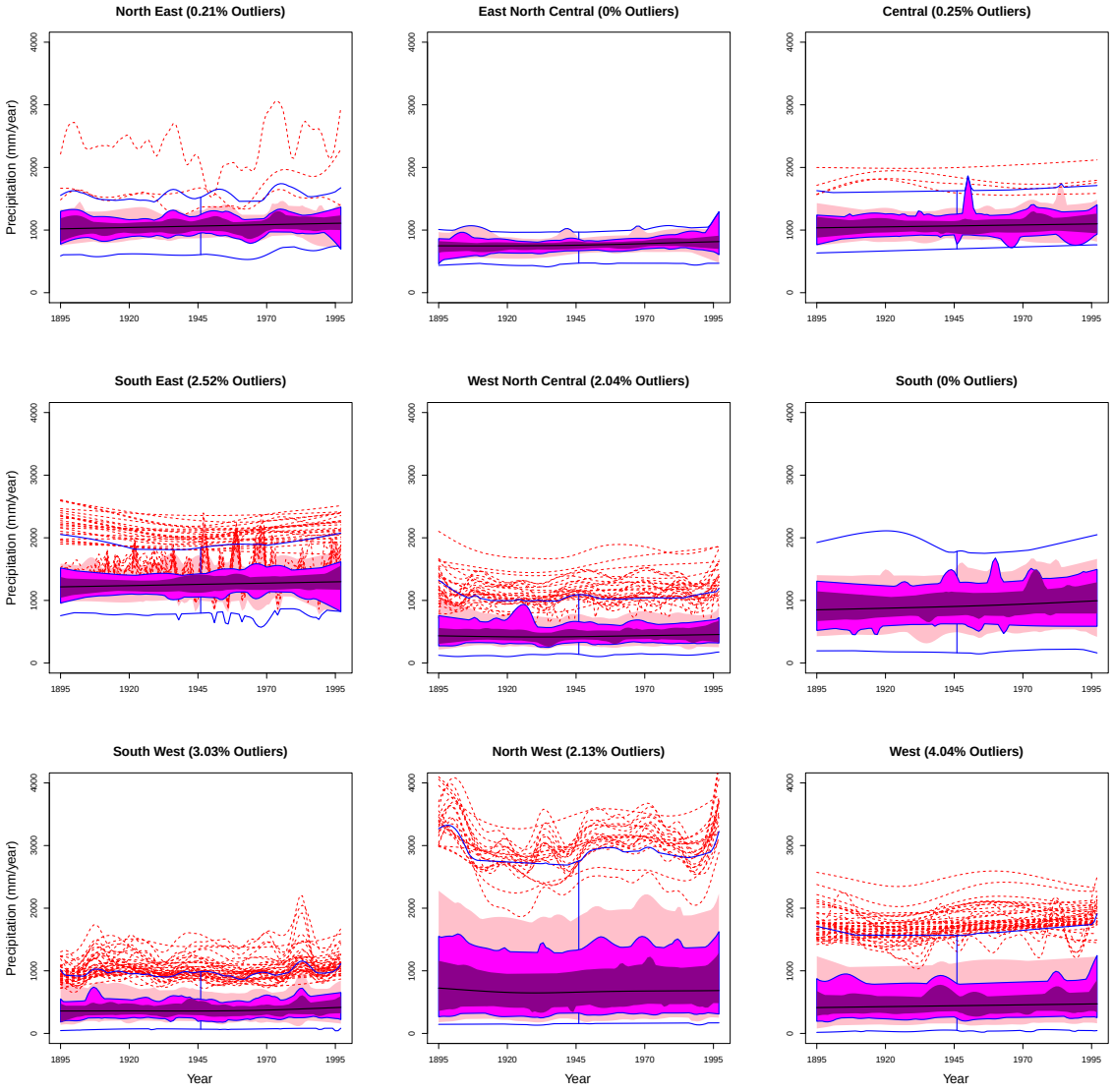

Figure 27. Enhanced functional boxplots of observed yearly precipitation over the nine climatic regions for the coterminous U.S. from 1895 to 1997 using MBD. Dark magenta, magenta and pink denote the 25%, 50% and 75% central regions respectively and the outlier rule is 1.5 times the 50% central region. The percentage of outliers in each climatic region is provided.

in the enhanced functional boxplot of North West in Figure 27. By identifying the locations of the potential outliers in the enhanced functional boxplot of South East, we notice that the annual precipitation at the southmost tip in Florida shows an oscillatory pattern. It varies greatly from year to year when hurricanes and droughts have occurred. In Florida, wet springs and summers make up the wet season, and relatively dry winters and autumns form the dry season. If we go back to look at the original monthly precipitation, it matches the wet and dry seasons at normal locations. However, the outlying locations usually have drier springs, but wet season from July to November even though it is during the drought. And during wet years, most of the precipitation is contributed by the period from July to November which is the hurricane season in Florida. Therefore, the high points of the oscillation in the enhanced functional boxplot capture the effects of hurricanes. If we use a logarithmic scale, it would yield fewer potential outliers. However, it is common that an observation could be an outlier in one scale but not in another. As the classical boxplot also suffers from the same problem, we prefer not to do any transformation in general.

As we have noticed, for spatio-temporal data, we do not have independent curves like in the children growth data example. These precipitation curves are spatially correlated, but the dependence between the curves should only affect the variance of the band depth estimator, not its unbiasedness. The percentage of potential outliers might be different because of the spatial correlation.

4.6. Discussion

This chapter presented the functional boxplot as an informative exploratory tool for visualizing functional data, as well as its generalization, the enhanced functional boxplot. These functional boxplots were applied to sea surface temperatures, children

growth and spatio-temporal precipitation datasets. With this new technique, outliers can be detected based on the 1.5 times the 50% central region empirical rule. Our approach is distinct from others in treating each curve as an observation rather than summarizing datasets pointwisely. The descriptive statistics in a functional boxplot are rank-based, hence they may lead to building robust statistical models to capture the features of complex datasets.

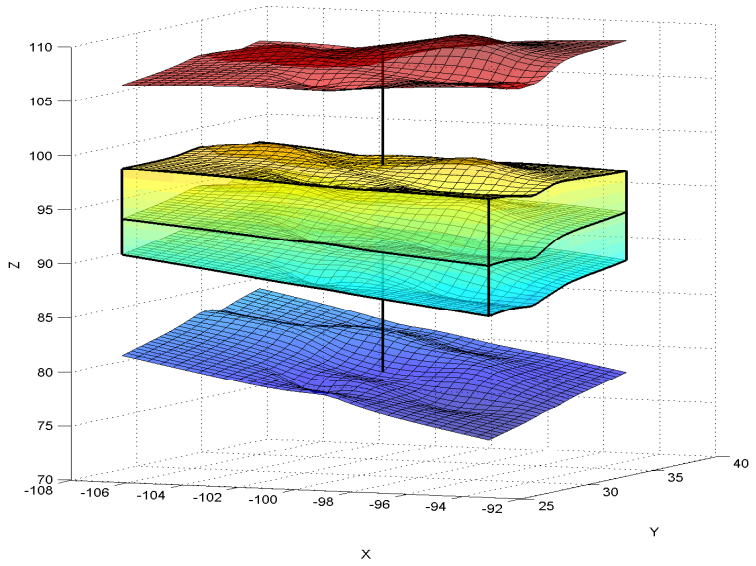

Figure 28. The surface boxplot with the box in the middle representing the 50% central region in $\mathbb{R}^3$, the middle surface inside the box denoting the median surface, and the upper and lower surfaces indicating the maximum non-outlying envelope.

For spatio-temporal data, we have viewed the information as a temporal curve at each spatial location. An alternative would be to treat the dataset as a spatial surface at each time. In that case, we could define a volume-based surface band depth for a surface S by counting the proportion of surface bands determined by J different surfaces ($2 \leq J \leq n$) in $\mathbb{R}^3$, containing S. This would lead to a three-dimensional

surface boxplot with similar characteristics as the functional boxplots defined in this article. An illustrative surface boxplot is shown in Figure 28. Similarly, the fences are obtained by the 1.5 times the 50% central region rule. Any surface outside the fences are flagged as outlier candidates. The surface boxplot is a natural extension of the functional boxplot to $\mathbb{R}^3$. However, to obtain a three-dimensional functional bagplot, one would definitely need robust principal component analysis techniques to an array rather than a matrix (Hyndman and Shang, 2010).

CHAPTER V

ADJUSTED FUNCTIONAL BOXPLOTS FOR SPATIO-TEMPORAL DATA VISUALIZATION AND OUTLIER DETECTION

5.1. Introduction

Functional data analysis is an attractive approach to study complex data in statistics. In many statistical experiments, the observations are functions by nature, such as temporal curves or spatial surfaces, where the basic unit of information is the entire observed function rather than a string of numbers. There is also an interesting class of applications that can be characterized as random processes evolving in space and time in, for instance, environmental science, agriculture, climatology, meteorology and hydrology.

To analyze functional data, many model-based methods have been developed over the years, among which Ramsay and Silverman (2005) provided various parametric methods while Ferraty and Vieu (2006) developed detailed nonparametric techniques. For spatio-temporal data, one can imagine a random field $Z(\mathbf{s}, t)$, $(\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$, observed at the space-time coordinates $(\mathbf{s}_1, t_1), \ldots, (\mathbf{s}_n, t_n)$. The spatio-temporal variable $Z(\mathbf{s}, t)$ could stand for temperature, precipitation, wind speed or atmospheric pollutant concentrations, to name a few. Some recent literature, such as Kyriakidis and Journel (1999), Brown et al. (2001), Banerjee et al. (2004), Schabenberger and Gotway (2005), and Cressie and Wikle (2011) point out the significance of the spatio-temporal modeling approach.

However, visualization methods can also help to display the data, highlight their characteristics and reveal interesting features. Sun and Genton (2011) proposed an

informative exploratory tool, the functional boxplot, and its generalization, the enhanced functional boxplot, for visualizing functional data as well as detecting potential outliers. The functional boxplot orders functional data by means of band depth (López-Pintado and Romo, 2009). It allows for ordering a sample of curves from the center outwards and, thus, introduces a measure to define the centrality or outlyingness of an observation. Indeed, one can compute the band depths of all the sample curves and order them according to decreasing depth values. Suppose each observation is a real function $y_i(t)$, $i = 1, \ldots, n$, $t \in \mathcal{I}$, where $\mathcal{I}$ is an interval in $\mathbb{R}$. Let $y_{[i]}(t)$ denote the sample curve associated with the ith largest band depth value. Then $y_{[1]}(t), \ldots, y_{[n]}(t)$ can be viewed as order statistics, with $y_{[1]}(t)$ being the deepest (most central) curve or simply the median curve, and $y_{[n]}(t)$ being the most outlying curve. The implication is that a smaller rank is associated with a more central position with respect to the sample curves. The order statistics induced by band depth start from the most central sample curve and move outwards in all directions. Thus, it is straightforward to define a central region for functional data.

In the classical boxplot, the box itself represents the middle 50% of the data. By analogy, the 50% central region in the functional boxplot can be defined by extending the concept of central region introduced by Liu et al. (1999) to functional data. The band delimited by the α proportion ($0 < \alpha < 1$) of deepest curves from the sample is used to estimate the α central region. In particular, the sample 50% central region is

$$C_{0.5} = \big\{(t, y(t)) : \min_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t) \leq y(t) \leq \max_{r=1,\ldots,\lceil n/2 \rceil} y_{[r]}(t)\big\},$$

where $\lceil n/2 \rceil$ is the smallest integer not less than $n/2$. The envelope of the 50% central region represents the box in a classical boxplot. Thus, this 50% central region is the analog to the "inter-quartile range" (IQR) and gives a useful indication of the spread of the central 50% of the curves.

For functional boxplots, based on the center outwards ordering induced by band depth for functional data, the descriptive statistics are: the envelope of the 50% central region, the median curve and the maximum non-outlying envelope. In addition, potential outliers can be detected in a functional boxplot by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. Recall that the 50% central region is the analog to the IQR. The outer region (the "fence") is obtained by inflating the inner region (the "envelope") by 1.5 times the range of the 50% central region. Any curves crossing the fences are flagged as potential outliers.

Considering that when each curve is simply a point, the functional boxplot degenerates to a classical boxplot, Sun and Genton (2011) suggested the constant factor 1.5 as in a classical boxplot, but left to the user the possibility of modifying it. However, for functional data, there will be necessarily dependence in time for each curve. And for spatio-temporal data, curves from different locations will be spatially correlated as well. The outlier detection performance may be affected by the dependence in time and space. Therefore, in this chapter, we investigate the relationship between the dependence and the constant factor, and then propose a method to adjust the factor in a functional boxplot. This leads to an adjusted functional boxplot. Febrero et al. (2007, 2008) also considered outlier detection in functional data by depth measures but they did not account for the temporal or spatio-temporal correlation in the data and their method is quite different from the functional boxplot approach.

Classical boxplots were first introduced by Tukey (1970) and Tukey (1977, pp. 39-43) in exploratory data analysis. In a classical boxplot, outliers can be detected by the 1.5 time IQR empirical rule. Here the constant factor 1.5 can be justified by a standard normal distribution. Let Q_1 and Q_3 be the first and third quartiles of the standard normal distribution, respectively. The fences determined by $L_1 = Q_1 - 1.5 \times IQR$ and $L_2 = Q_3 + 1.5 \times IQR$ are -2.698 and 2.698. Then the probability of being detected

as an outlier is 0.7%. If we change the factor to 2, then the probability that a value is an outlier is only 0.07%. Therefore, in the functional boxplot, we would like to adjust the value of the constant factor based on the dependence such that the probability of detecting no outliers is 99.3% when actually no outliers are present. It is clear that in a functional boxplot, the factor adjustment is crucial for outlier detection since it determines the percentage of detected outliers. However, the adjustment involves a certain amount of computation, thus it is not necessary if one only wants to visualize and compare functional or spatio-temporal data.

This chapter is organized as follows. Section 5.2 illustrates how the dependence in time and space affects the outlier detection performance of functional boxplots. Then the new method to select the constant factor in a functional boxplot is proposed in Section 5.3. The adjusted functional boxplots are demonstrated on applications to space-time datasets in Section 5.4, and a discussion is provided in Section 5.5.

5.2. Simulation Studies

To illustrate how the dependence in time and space affects the outlier detection performance of the functional boxplots, simulation studies are conducted under different spatio-temporal covariance models reviewed by Gneiting et al. (2007). Other covariance models can be found in Cressie and Huang (1999) and Gneiting (2002).

5.2.1. Data generation

We consider data drawn from a zero-mean, stationary spatio-temporal Gaussian random field $Z(\mathbf{s}, t)$ where $(\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}$. Let $C(\mathbf{h}, u) = \text{cov}\{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)\}$ be the covariance function between any two observations whose locations are apart by a vector $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$ and a time span $u = |t_1 - t_2|$. Then $C(\mathbf{h}, 0)$ and $C(\mathbf{0}, u)$ are purely

spatial and purely temporal covariance functions, respectively. The spatio-temporal data $Z(\mathbf{s}_i, t)$, where $1 \leq i \leq n$, $t \in [0, 1]$, are generated from the Gaussian random field $Z(\mathbf{s}, t)$ at $n = 100$ locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ on a grid of size 10×10 with the grid spacing $1/9$. We aim at seeing how the strength of the correlation in time, in space, or in both, affects the outlier detection performance of the functional boxplot with the constant factor $F = 1.5$. We consider the following isotropic correlation models:

1. A purely temporal correlation function of Cauchy type,

$$C_T(u) = (1 + a|u|^{2\alpha})^{-1}, \tag{5.1}$$

   where $\alpha \in (0, 1]$ controls the strength of the temporal correlation and $a > 0$ is the scale parameter in time. We set $a = 1$ and let α vary from 0.1 to 0.9.

2. A purely spatial correlation function of the form

$$C_S(\mathbf{h}) = (1 - \nu) \exp(-c\|\mathbf{h}\|) + \nu\delta_{\mathbf{h}=\mathbf{0}}, \tag{5.2}$$

   where $c > 0$ controls the strength of the spatial correlation, and $\nu \in (0, 1]$ is a nugget effect. We set $\nu = 0.05$ and let c vary from 0.1 to 2.

3. A space-time separable correlation function of the form

$$C_{SEP}(\mathbf{h}, u) = C_S(\mathbf{h})C_T(u), \tag{5.3}$$

   which is the product of the purely temporal correlation function (5.1) and the purely spatial correlation function (5.2). Here, we consider combinations of α and c, where each has three levels, $\alpha = 0.1, 0.5, 0.9$ and $c = 0.1, 1, 2$.

4. A fully symmetric but generally non-separable correlation function

$$C_{FS}(\mathbf{h}, u) = \frac{1 - \nu}{1 + a|u|^{2\alpha}} \left[\exp\left\{ -\frac{c\|\mathbf{h}\|}{(1 + a|u|^{2\alpha})^{\beta/2}} \right\} + \frac{\nu}{1 - \nu}\delta_{\mathbf{h}=\mathbf{0}} \right],$$

where $0 \leq \beta \leq 1$ controls the non-separability. It reduces to the separable model (5.3) when $\beta = 0$. Here we set $\beta = 1$, the most non-separable version of this model, $\nu = 0.05$ and consider the same combinations of α and c as in model (5.3).

5. A general stationary correlation model

$$C_{STAT}(\mathbf{h}, u) = \frac{(1-\nu)(1-\lambda)}{1+a|u|^{2\alpha}} \left[\exp\left\{ -\frac{c\|\mathbf{h}\|}{(1+a|u|^{2\alpha})^{\beta/2}} \right\} + \frac{\nu}{1-\nu} \delta_{\mathbf{h}=\mathbf{0}} \right] + \lambda \left(1 - \frac{1}{2v} |h_1 - vu| \right)_+ ,$$

where $q_+ = \max(q, 0)$ and h_1 is the first component of the spatial separation vector $\mathbf{h} = (h_1, h_2)'$. Here $0 \leq \lambda \leq 1$ controls the asymmetry. Again, we set $a = 1$, $\beta = 1$ and $\nu = 0.05$, then let $\lambda = 0.5$, $v = 0.05$ and consider the same combinations of α and c as in model (5.3).

In the simulation studies, we generate $n = 100$ curves without any outliers at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and $p = 50$ time points from the model $Z(\mathbf{s}, t) = g(\mathbf{s}, t) + e(\mathbf{s}, t)$, with mean $g(\mathbf{s}, t) = 0$, $(\mathbf{s}, t) \in [0, 1]^2 \times [0, 1]$ and where $e(\mathbf{s}, t)$ is a Gaussian random field with zero mean and covariance function the same as each of the correlation models above. Then with 1,000 replications, we compute the proportion of time that functional boxplots with the constant factor $F = 1.5$ detect no outliers. Thus, a proportion much smaller than 1 is an indication of bad outlier detection performance.

5.2.2. Numerical results

Tables 7 and 8 summarize the simulation results. In the purely temporal model (5.1), the larger the value of α, the stronger the temporal dependence is when $u < 1$. For a fixed constant factor $F = 1.5$ in the functional boxplot, Table 7 shows that the proportion of times that the functional boxplot correctly detects no outliers decreases

as α increases. In other words, the stronger the correlation in time, the worse the outlier detection performance is. Similarly, in the purely spatial model (5.2), the smaller the value of c, the stronger the spatial dependence is. For all the values of c in Table 7, the proportions of correctly detecting no outliers are close to 1. This is an evidence that the constant factor 1.5 is too large when spatial correlation exists since usually spatially correlated curves are more concentrated than independent ones.

Table 7: The proportion of times (p) that a functional boxplot with the constant factor $F = 1.5$ correctly detects no outliers under the purely temporal and the purely spatial correlation models with 1,000 replications and $n = 100$ curves.

Temporal	α	0.1	0.3	0.5	0.7	0.9
	p	0.998	0.974	0.938	0.839	0.745
Spatial	c	0.1	0.5	1	1.5	2
	p	1	0.999	1	1	1

Table 8: The proportion of times that a functional boxplot with the constant factor $F = 1.5$ correctly detects no outliers under the separable, symmetric but non-separable and the general stationary spatial-temporal correlation models with 1,000 replications and $n = 100$ curves.

c / α	Separable			Symmetric			Stationary		
	0.1	1	2	0.1	1	2	0.1	1	2
0.1	1	0.997	0.993	0.995	0.990	0.994	0.995	0.994	0.995
0.5	1	0.980	0.961	0.956	0.945	0.952	0.957	0.943	0.950
0.9	1	0.942	0.908	0.901	0.854	0.836	0.900	0.833	0.844

Table 8 provides the proportion of times that a functional boxplot with the constant factor $F = 1.5$ correctly detects no outliers for each combination of α and c under the separable, symmetric but non-separable and the general stationary spatial-temporal correlation models. It also shows that the proportion of correctly detecting no outliers decreases as α or the temporal dependence increases. For each value of α under different correlation models, since the strongest spatial correlation ($c = 0.1$) makes curves most concentrated, with the fixed constant factor $F = 1.5$ in functional boxplots, the proportions of correctly detecting no outliers for $c = 0.1$ are always the largest among those for $c = 0.1, 1, 2$. When the dependence in time is relatively large, $\alpha = 0.5, 0.9$, all the proportions under the separable correlation model are greater, hence better outlier detection performance, than that under either the symmetric but non-separable or the general stationary correlation model. This suggests that the interaction or the separability between spatial and temporal dependence has an effect on the outlier detection performance in a functional boxplot. To investigate the possible effect of the asymmetry in a correlation model, we compare the proportion of a functional boxplot correctly detecting no outliers under the symmetric but non-separable and the general stationary correlation models. When $\alpha = 0.5, 0.9$ and $c = 1$, the two proportions under the symmetric correlation model are larger than those under the general stationary one. However, there are still several cases where the general stationary model shows a better outlier detection performance.

It is now clear that the adjustment of the constant factor in functional boxplots is necessary when spatio-temporal correlations exist. In Section 5.3, we propose a method for selecting the factor F.

5.3. Selection of the Adjustment Factor

The simulation studies in Section 5.2 have shown that the constant factor $F = 1.5$ gives different probability coverage under different spatio-temporal correlation models. In other words, we should choose the value of the factor F by controlling the probability of detecting no outliers to be 99.3% when actually no outliers are present.

Sharing the same idea as in the simulations, we propose first to estimate the covariance matrix of the data in order to generate observations without any outliers. Then we use the simulations described in Section 5.2 to choose the constant factor F such that the percentage of outliers in a functional boxplot $1 - 99.3\% = 0.7\%$. Finally, we can apply this adjusted factor to the functional boxplot on the original data and detect outliers. When estimating the covariance matrix, robust techniques are needed since outliers may exist in the original data. We use a componentwise estimator of a dispersion matrix proposed by Ma and Genton (2001), based on a highly robust estimator of scale, Q_n. This estimator is location-free and has already been successfully used in the context of variogram estimation (Genton, 1998) in spatial statistics, and autocovariance estimation (Ma and Genton, 2000) in time series. There are also many other robust estimators that could be used, some of which are based on the minimization of a robust scale of Mahalanobis distances. For example, the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators (Rousseeuw, 1984, 1985). However, their computation can be challenging. Alternatively, a more rapid orthogonalized Gnanadesikan-Kettenring (OGK) estimator was proposed by Maronna and Zamar (2002) for high dimensional datasets.

In order to reduce the computational burden and simplify the covariance matrix estimation, we only generate a small number of curves, $n = 100$, without any outliers at p time points from the model $Z(\mathbf{s}, t) = g(\mathbf{s}, t)+e(\mathbf{s}, t)$, with mean $g(\mathbf{s}, t) = 0$, $(\mathbf{s}, t) \in$

$\mathbb{R}^2 \times \mathbb{R}$. Here $e(\mathbf{s}, t)$ is a Gaussian random field with mean zero and covariance function estimated from the standardized original data, hence with marginal variance 1. For simplicity, we let the trend $g(\mathbf{s}, t)$ be 0 and the marginal variance be 1 because they do not affect the values of band depth, hence the order of these curves. In addition, for spatio-temporal data, to reduce the dimension of the spatio-temporal covariance, we only estimate the covariances at certain distances and time lags depending on the simulation design.

5.4. Applications

5.4.1. Sea surface temperatures

A dataset of sea surface temperatures was used by Hyndman and Shang (2010) and Sun and Genton (2011) to demonstrate the functional bagplot and the functional boxplot, respectively. The data consist of monthly sea surface temperatures (SST) measured in degrees Celsius over the east-central tropical Pacific Ocean and are shown in the left panel of Figure 29. In this case, each curve represents one year of observed SST in degrees Celsius from January 1951 to December 2007. The functional boxplot with the constant factor $F = 1.5$ in Sun and Genton (2011) detects two potential outliers: the years 1983 and 1997. In addition, the year 1982 from September to December and the year 1998 from January to June are viewed as being part of the maximum envelope. These 57 annual temperature curves show similarity in shape since they share a common mean function. Therefore, we detrend them first by subtracting the sample mean at each time point and then check the correlations between the curves. Since the correlations are not statistically significant, we assume that these annual temperature curves are independent copies of each other and estimate the 12×12 covariance matrix in time. In simulations, by generating $n = 100$ curves at

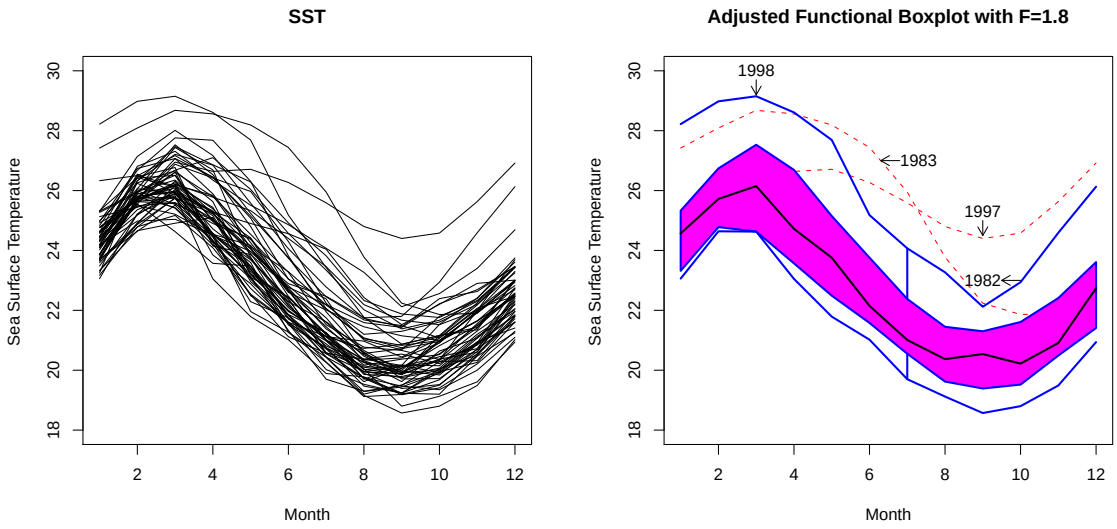

Figure 29. Left panel: data of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean from 1951 to 2007. Right panel: the adjusted functional boxplot of SST with the constant factor 1.8. The blue lines denote envelopes and the black line represents the median curve. The red dashed curves are the outlier candidates detected by the 1.8 times the 50% central region rule.

$p = 12$ time points from a Gaussian process with zero mean and estimated covariance function, the coverage probabilities for different values of the constant factor are listed in Table 9 with 1,000 replications. We select the constant factor to be 1.8 since when $F = 1.8$, the coverage probability is 0.995 close to 99.3%. The adjusted functional boxplot of the sea surface temperatures with the constant factor 1.8 is shown in the right panel of Figure 29. After adjusting the constant factor, the functional boxplot still detects two El Niño years as outlier candidates.

Table 9: The coverage probabilities for different values of the constant factor F with $n = 100$ curves at $p = 12$ time points and 1,000 replications in simulations for the sea surface temperatures example. The selected factor is in bold font.

Factor F	1.2	1.3	1.4	1.5	1.6	1.7	**1.8**	1.9	2.0
Coverage	0.768	0.859	0.922	0.956	0.979	0.988	**0.995**	1.000	1.000

5.4.2. Spatio-temporal precipitation

The functional boxplot can summarize information from complex data, such as space-time datasets. Sun and Genton (2011) illustrated this aspect by visualizing the observed annual total precipitation data for the coterminous U.S. from 1895 to 1997, provided by the Institute for Mathematics Applied to Geosciences at the web page (*http://www.image.ucar.edu/Data/US.monthly.met/*). There are 11,918 stations reporting precipitation at some time in this period. The observations are time series with $p = 103$ yearly precipitation observations, or one curve, at each spatial location. Functional boxplots were applied to nine climatic regions for precipitation in the U.S. defined by the National Climatic Data Center (NCDC) and the percentages of detected potential outliers for each region were reported. Sun and Genton (2011) noticed that for spatio-temporal data, the precipitation curves are not independent but spatially correlated. Therefore, the percentages of potential outliers might not be correct due to the spatial correlations.

By taking the spatio-temporal correlation into account, we adjust the constant factor in the functional boxplots again by simulations. For each climatic region, in the simulation, we generate spatio-temporal data from a zero-mean Gaussian random field at $n = 100$ locations on a grid of size 10×10 with the grid spacing $1/9$ at $p = 30$ time points. Here the distance unit is kilometer and the time unit is year. Then

Table 10: The coverage probabilities for different values of the constant factor $F = 1.4, 1.5, \ldots, 2.2$ with $n = 100$ curves at $p = 30$ time points and 1,000 replications in simulations for the precipitation application. The nine climatic regions are North East (NE), East North Central (ENC), Central (C), South East (SE), West North Central (WNC), South (S), South West (SW), North West (NW), and West (W). The selected factors are in bold font.

Region	1.4	1.5	1.6	1.7	**1.8**	**1.9**	**2.0**	**2.1**	2.2
NE	0.927	0.951	0.971	0.985	0.987	**0.993**	0.994	0.995	0.996
ENC	0.896	0.941	0.958	0.972	0.985	0.989	**0.993**	0.994	0.995
C	0.926	0.951	0.970	0.984	0.988	**0.994**	0.995	1.000	1.000
SE	0.926	0.948	0.974	0.979	0.988	**0.993**	0.996	0.999	1.000
WNC	0.893	0.944	0.966	0.973	0.984	0.990	**0.992**	0.995	0.996
S	0.902	0.934	0.959	0.976	0.983	0.987	0.989	**0.993**	0.994
SW	0.920	0.948	0.972	0.979	0.987	0.992	**0.993**	0.995	0.995
NW	0.911	0.939	0.962	0.974	0.983	0.987	0.992	**0.993**	0.995
W	0.911	0.949	0.974	0.988	**0.993**	0.994	0.996	0.999	0.999

we estimate the $3,000 \times 3,000$ covariance matrix from the standardized original data and obtain the coverage probabilities for different values of the constant factor with 1,000 replications. The results are summarized for each region in Table 10. Each component of the $3,000 \times 3,000$ covariance matrix is estimated by the Q_n-based procedure of Ma and Genton (2001). To reduce the computational effort, for each combination of time lag and distance, we estimate the covariance element by randomly selecting pairs of the irregularly spaced locations that are close to the distances on the 10×10 grid under the assumption of stationarity. Based on the concept of central regions, Sun and Genton (2011) generalized the functional boxplot to an

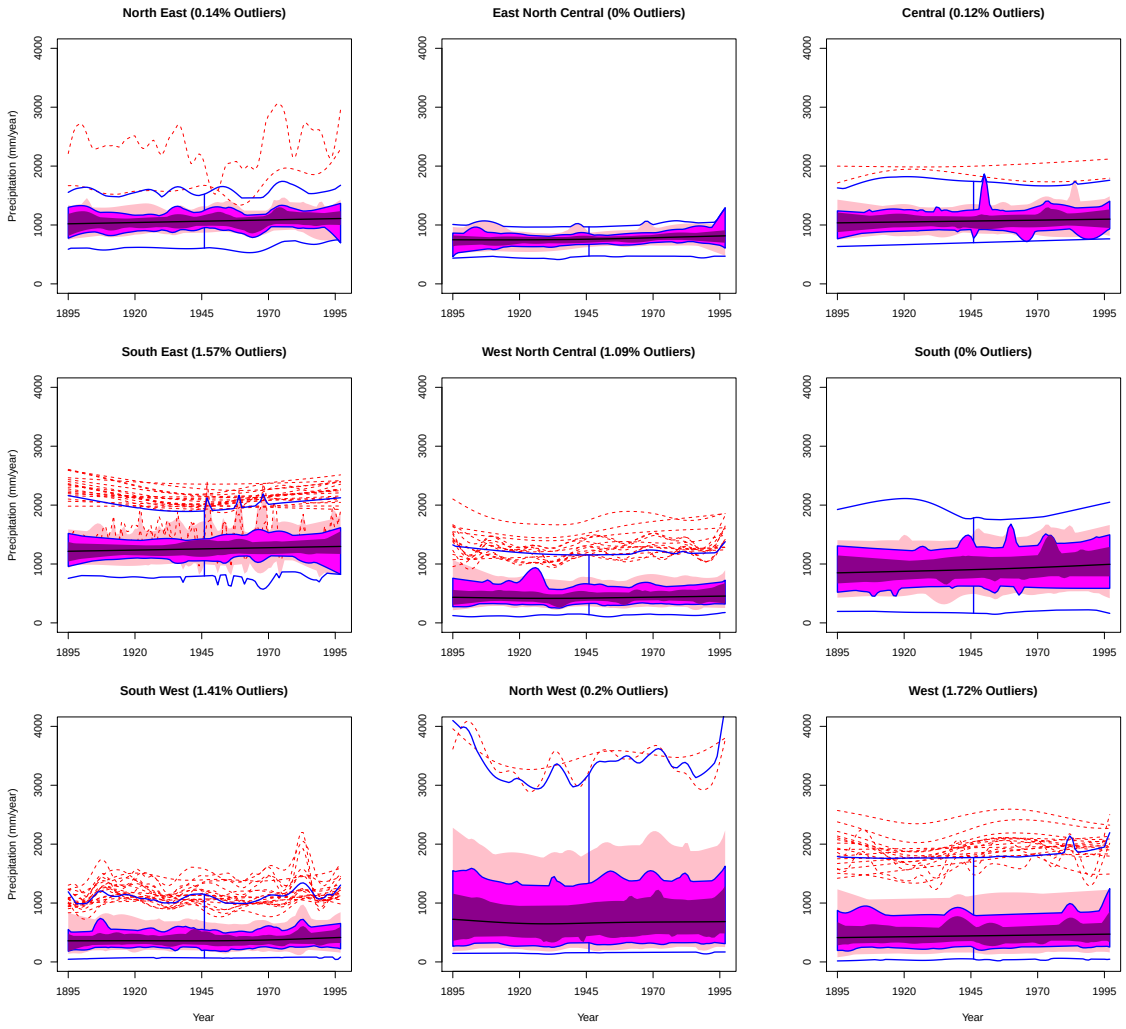

Figure 30. Adjusted enhanced functional boxplots of observed yearly precipitation over the nine climatic regions for the coterminous U.S. from 1895 to 1997. Dark magenta, magenta and pink denote the 25%, 50% and 75% central regions, respectively, and the outlier rule is the adjusted constant factor times the 50% central region. The percentage of detected outliers in each climatic region is provided.

enhanced functional boxplot where the 25% and 75% central regions are provided in addition to the 50% central region. For the spatio-temporal precipitation data, the nine adjusted enhanced functional boxplots in Figure 30 can still reveal information about the different annual precipitation characteristics for different climatic regions, but with less potential outliers than previously detected by Sun and Genton (2011). The percentage of detected outliers for each region is summarized in Table 11.

Table 11: Comparison of outlier detection percentages for each climatic region before and after adjustment of the constant factor.

Region	NE	ENC	C	SE	WNC	S	SW	NW	W
Before	0.21	0	0.25	2.52	2.04	0	3.03	2.13	4.04
After	0.14	0	0.12	1.57	1.09	0	1.41	0.20	1.72

5.4.3. General Circulation Model

A General Circulation Model (GCM) is a climate model of the general circulation of a planetary atmosphere or ocean. It uses complex computer programs to simulate the Earth's climate system and allows us to look into the Earth's past, present and future climate states. Here we consider precipitation data generated from the National Center for Atmospheric Research-Community Climate System Model (CCSM) Version 3.0 (Collins et al., 2006, and references therein), which was run given scenarios from the Intergovernmental Panel on Climate Change (IPCC)'s Special Report on Emission Scenarios (SRES); see IPCC (2000) and Ammann et al. (2010). Functional boxplots can be used to visually compare the annual precipitation produced by the GCM with the real observations from weather stations considered in the previous

section.

For these GCM data, there are 256×128 cells over the whole globe with a resolution of 1.406×1.406 degree, or around 156×156 kilometers. The observations from weather stations are much denser but only for the coterminous U.S.. To make the weather station observations comparable to the GCM data, we match them by longitude and latitude first, and then average the observations from weather stations within each cell which leads to 473 cells in total, hence 473 annual precipitation curves for the coterminous U.S..

For the coterminous U.S. precipitation, the functional boxplots with the constant factor $F = 1.5$ for weather station and GCM data are shown in the top panel of Figure 31 with the percentage of detected outliers. Now, we estimate the $3,000 \times 3,000$ spatio-temporal covariance matrix from the standardized original data by the same simulation design as in Section 5.4.2 and the coverage probabilities for different values of the constant factor with 1,000 replications are summarized in Table 12 for both weather station and GCM data. The adjusted functional boxplots with percentage of detected outliers are shown in the bottom panel of Figure 31.

Table 12: The coverage probabilities for different values of the constant factor $F = 1.4, 1.5, \ldots, 2.2$ with $n = 100$ curves at $p = 30$ time points and 1,000 replications in simulations for both weather station and GCM data. The weather station and the GCM past are for the time period from 1970 to 1997. The GCM future is for the time period from 2070 to 2097. The selected factors are in bold font.

Source	1.4	1.5	1.6	1.7	1.8	**1.9**	**2.0**	**2.1**	2.2
Weather Stations	0.904	0.945	0.968	0.979	0.987	**0.993**	0.997	0.998	0.999
GCM Past	0.914	0.950	0.971	0.979	0.985	0.990	0.992	**0.993**	0.996
GCM Future	0.912	0.947	0.966	0.981	0.988	0.991	**0.994**	0.998	0.998

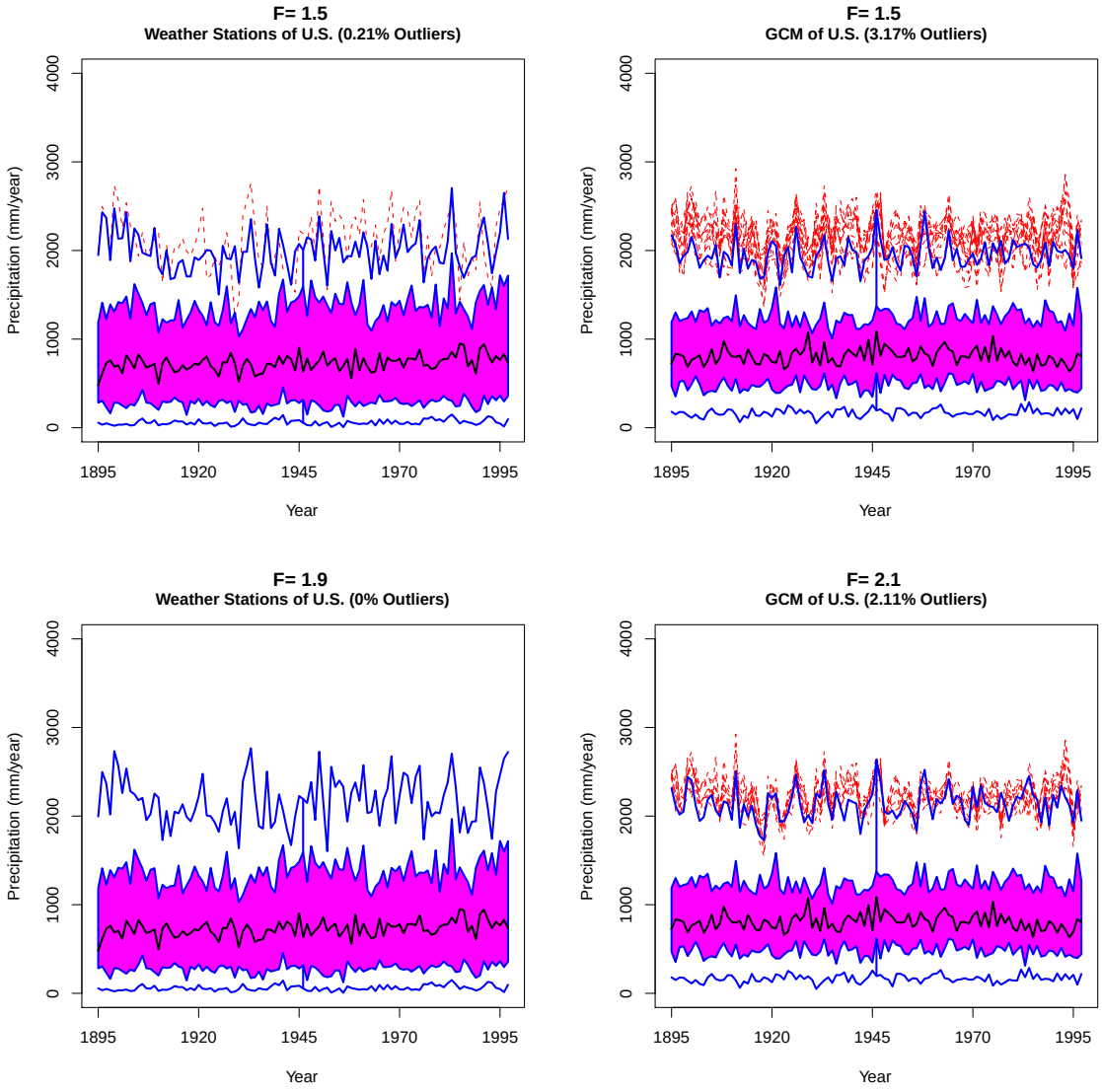

Figure 31. Top panel: the functional boxplots of weather station and GCM data with the constant factor $F = 1.5$ for the coterminous U.S. precipitation. Bottom panel: the adjusted functional boxplots of weather station and GCM data with the adjusted constant factor F for the coterminous U.S. precipitation.

Figure 31 shows that the precipitation data produced by the GCM roughly capture the overall patten of the U.S. precipitation. The two functional boxplots of weather station and GCM data coincide on the median curves and the maximum annual precipitation for some wet locations is also on the same level. However, the narrower 50% central region in the functional boxplot of the GCM data indicates the first 50% most representative precipitation curves have less variability which leads to a relatively large percentage of outliers. Moreover, both functional boxplots are skewed to the right (i.e. to large precipitation), but the one for GCM does not produce as low annual precipitation as the real observations from weather stations for some dry locations.

The maps of weather station and GCM data where outliers are detected by the adjusted functional boxplots with respect to either the whole U.S. or each of the climatic regions are shown in Figure 32. For the whole U.S., the outliers, denoted by red plus signs, are around the North of the U.S. for GCM data, but the only one outlier detected by the functional boxplot is located at the North West for weather station data. The maps also show that the precipitation data produced by GCM do not capture the characteristics of the observed precipitation from weather stations well. From weather stations, the West of the U.S. overall has a lower precipitation than the East, and the higher precipitation locations are along the west coast and the South East. This pattern is hard to see from the GCM and the higher precipitation locations appear in the North shown as outliers. As can be expected, the detected outliers with respect to each climatic region are different from those for the whole U.S.. For climatic regions, the potential outliers are still in the North of the U.S. for GCM data with a larger percentage, but are located along the West coast and the Rocky Mountain area for weather station data.

The GCM also simulates precipitation for the future. The future runs of the GCM

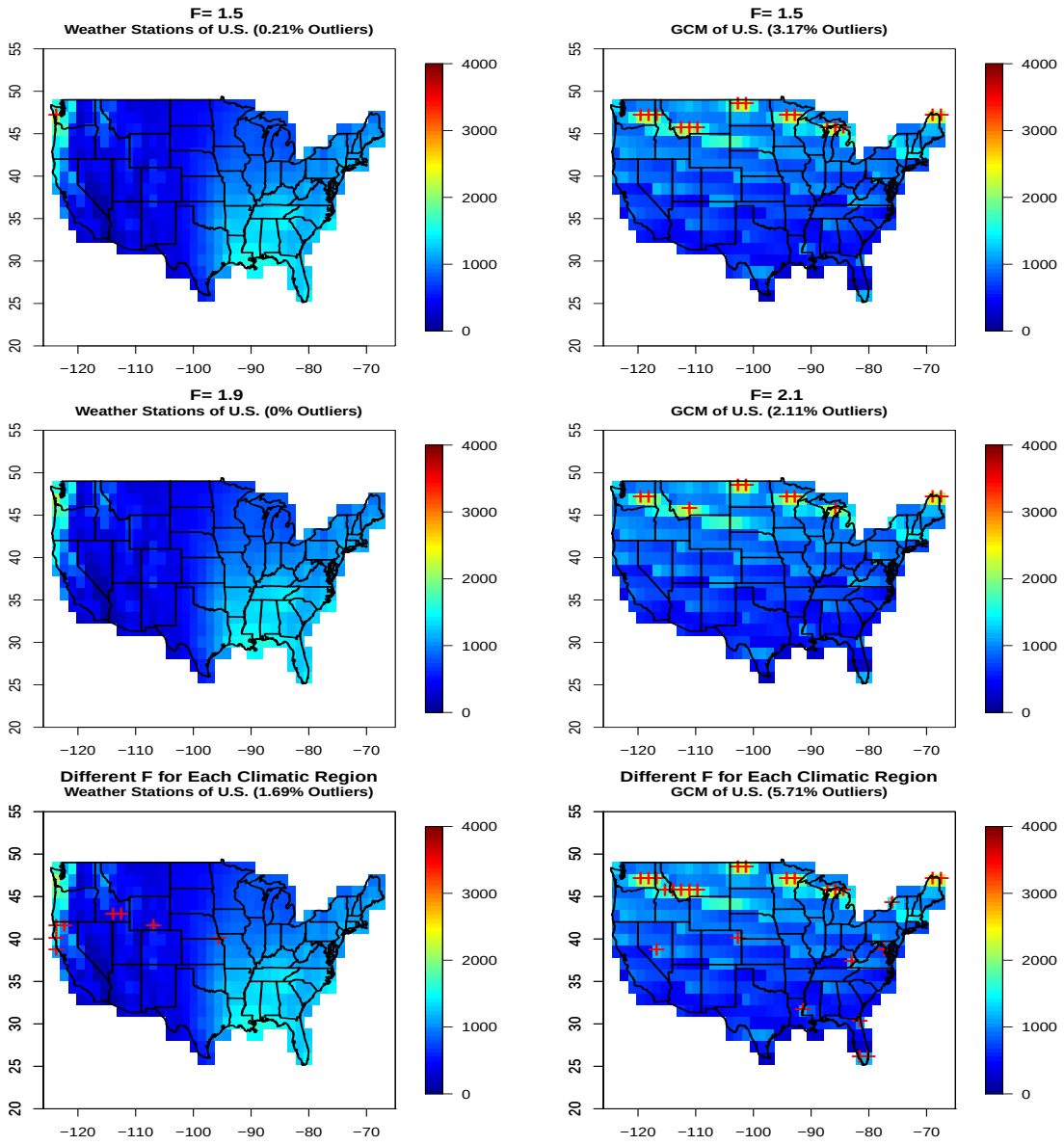

Figure 32. Top panel: the maps of weather station and GCM data where outliers are detected by the functional boxplots with the constant factor $F = 1.5$ for the coterminous U.S. precipitation. Middle panel: the maps of weather station and GCM data where outliers are detected by the adjusted functional boxplots for the coterminous U.S. precipitation. Bottom panel: the maps of weather station and GCM data where outliers are detected by the adjusted functional boxplots for each climatic region. The colors of each cell denote the averaged annual precipitation and the red plus signs indicate the detected outliers.

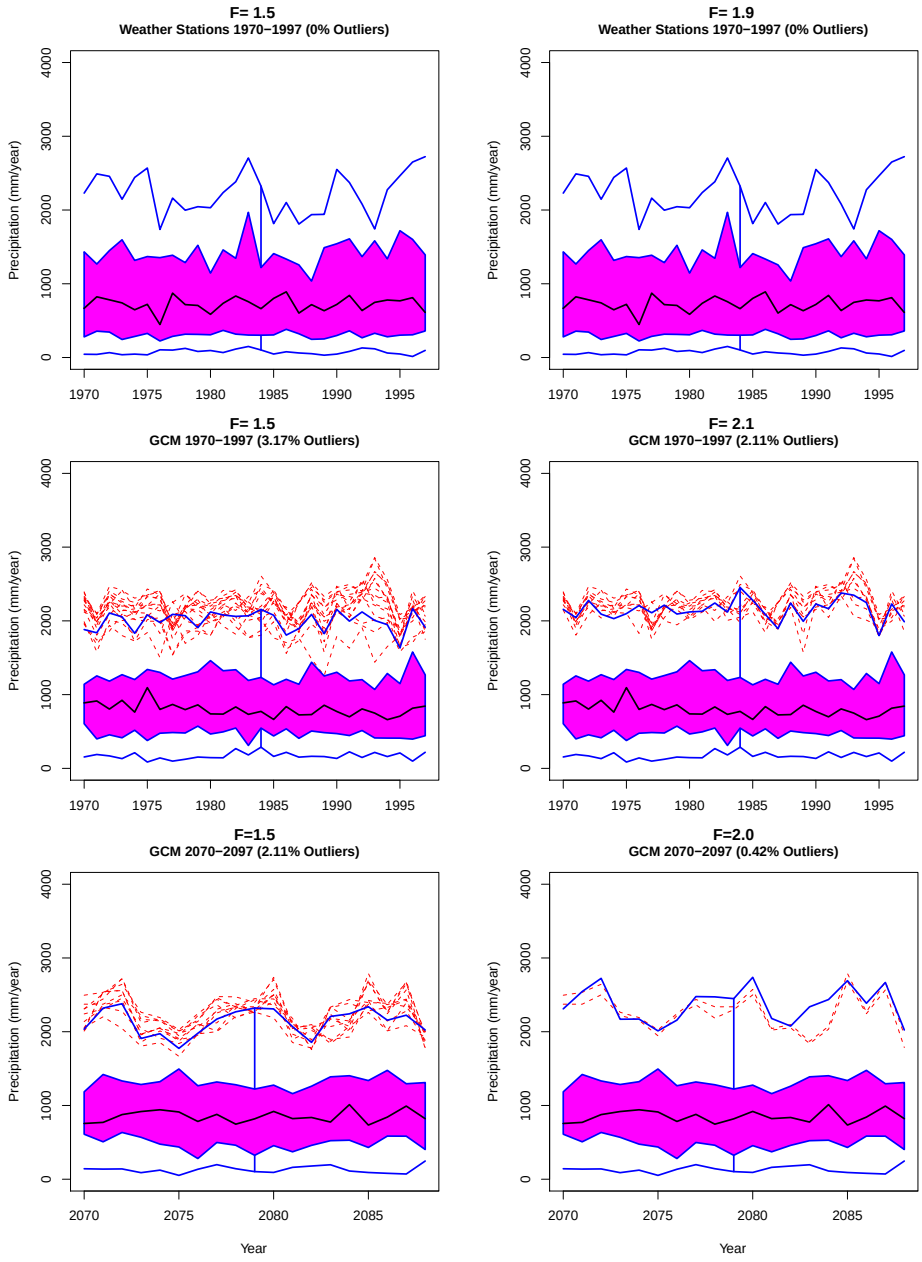

Figure 33. Top panel: the functional boxplot and the adjusted functional boxplot of weather station data for the coterminous U.S. precipitation from 1970 to 1997. Middle panel: the functional boxplot and the adjusted functional boxplot of GCM data for the coterminous U.S. precipitation from 1970 to 1997. Bottom panel: the functional boxplot and the adjusted functional boxplot of GCM data for the coterminous U.S. precipitation from 2070 to 2097 under IPCC A2 scenario.

were under the IPCC A2 scenario (Ammann et al. 2010) after the year 2020, which considers a continued increase of atmospheric green house gases and the associated warming throughout the 21st century. To compare the past precipitation with the future rainfall, we use the adjusted functional boxplots to visualize the spatio-temporal precipitation for the time period from 1970 to 1997 and for the future period from 2070 to 2097. For the past, the functional boxplots of weather station and GCM data are shown in the top and middle panels of Figure 33. The bottom panel of Figure 33 shows the functional boxplots of GCM data for the future. The future runs of the GCM produce a little wider 50% central region than the past runs do, thus a smaller outlier percentage, but both have narrower 50% central regions hence larger outlier percentages compared to the weather station data. We can also see that the median curves of the past and future runs from GCM are higher than that from weather stations and the minimum precipitation is also higher than the real observations. The corresponding maps are shown in Figure 34 including the outliers with respect to each of the climatic regions. The map of the GCM future runs follows the same pattern with the GCM past runs which is different from the weather stations. For the whole U.S., the detected outliers for the GCM future runs are fewer than those for the past runs. However, for the climatic regions, the detected outlier percentage by the GCM future runs is closer to the percentage from the past runs with less outliers in the West but more outliers in the East.

5.5. Discussion

This chapter has focused on how to adjust the functional boxplot proposed by Sun and Genton (2011) for correlations in order to perform functional and spatio-temporal data visualization and outlier detection. In a functional boxplot, potential outliers

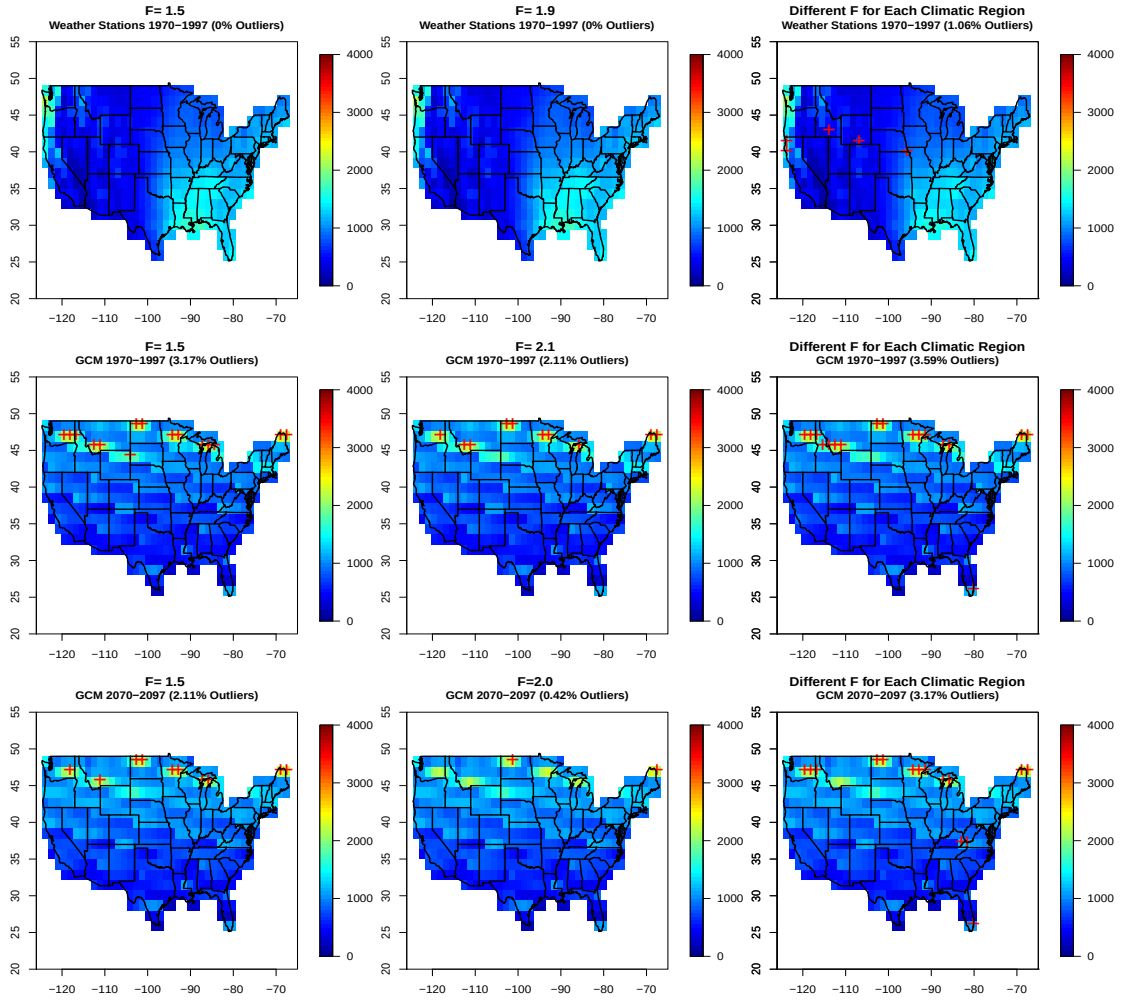

Figure 34. Top left and top center panels: the maps of weather station data where outliers are detected by the functional boxplot with the constant factor $F = 1.5$ (left) and the adjusted functional boxplot (center) for the coterminous U.S. precipitation from 1970 to 1997. Top right panel: the map of weather station data where outliers are detected by the adjusted functional boxplots for each climatic region for the time period from 1970 to 1997. Middle panels: the three maps of GCM data corresponding to the top panels for the time period from 1970 to 1997. Bottom panels: the three maps of GCM data corresponding to the top panels for the time period from 2070 to 2097 under IPCC A2 scenario. The colors of each cell denote the averaged annual precipitation and the red plus signs indicate the detected outliers.

can be detected by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. However, for functional data, there is necessarily dependence in time for each curve. And for spatio-temporal data, curves from different locations are spatially correlated as well. The simulation studies in Section 5.2 showed that the outlier detection performance is obviously affected by the dependence in time and space. Therefore, to correct the outlier detection performance, the constant factor of the empirical outlier rule is important. The factor $F = 1.5$ in a classical boxplot can be justified by a standard normal distribution, since it leads to a probability of 0.993 that any given observation is not an outlier. Following this idea, we proposed a simulation-based method to select this constant factor for a functional boxplot by controlling the percentage of non-outliers to be 99.3% when actually no outliers are present. Then how to estimate the covariance function, especially for spatio-temporal data, is also important and robust techniques are needed when considering the potential presence of outliers in the original data.

As applications, we used our method to adjust the functional boxplots for sea surface temperatures, spatio-temporal precipitation and GCM data. In fact, all the selected factors were greater than 1.5 which agrees with the simulation results in Section 5.2. The interpretation is that a positive correlation leads to larger variability, therefore, the extreme observations may not be outliers but may be due to the positive correlation in time and space.

For spatio-temporal data, we have viewed the information as a temporal curve at each spatial location. Sun and Genton (2011) also proposed an alternative to treat such data as a spatial surface at each time. In this case, it would lead to a three-dimensional surface boxplot with similar characteristics as the functional boxplots. Similarly, for outlier detection, the fences are obtained by the 1.5 times the 50% central region rule. Any surfaces crossing the fences are flagged as outlier candidates.

Therefore, for the surface functional boxplot in $\mathbb{R}^3$, the constant factor can also be adjusted by the simulation-based method described in this chapter and leads to an adjusted surface functional boxplot.

CHAPTER VI

CONCLUSIONS

In this dissertation, we have discussed the inference and visualization of periodic sequences. Specifically, we have proposed a cross-validation period estimator for equally spaced data. The method is computationally simple and implicitly penalizes multiples of the smallest period. Given a particular period, or cycle length, a leave-out-one-cycle version of CV is used to compute an average squared prediction error. The cycle length minimizing this average squared error is the period estimator. It is shown both theoretically and by simulation that the CV method has a much higher probability of choosing the correct model than it does in familiar cases where the considered models are nested. Our theory shows that the CV period estimator $\hat{p}$ is virtually consistent for large p, in that its asymptotic probability of equaling p increases monotonically to 1 as p becomes large. When $p = 15$ this probability is approximately 0.99. It is worth noting that the CV method has the advantage that it can easily deal with missing data, as long as the missing data are at random. Moreover, models corresponding to different periods may be ranked from best to worst by considering values of the objective function, thus, extending the possibilities of interpretation.

As a multivariate extension, we have also proposed a cross-validation period estimator for multiple equally spaced periodic sequences. Sharing a similar idea with the CV method for one sequence, a leave-out-one-cycle version of CV is used to compute an average squared prediction error given a particular period, or cycle length. The multivariate CV method uses the conditional means, i.e., conditional on other correlated sequences, to predict the left out cycle in cross-validation. In this way, the period estimation for a sequence X has been improved by borrowing information from

other sequences with which X is correlated. In theory, we show that the asymptotic behavior of the bivariate CV is the same as the CV for one sequence. In our simulation studies, however, it is shown that for finite samples, the better the periods of the other correlated sequences are estimated, the more substantial improvements can be obtained in estimating the period of interest. We have also shown that more correlated sequences lead to more improvements, although the improvement from 1 to 2 sequences is apparently larger than improvements from 2 to 3, etc. In addition to the CV method, we also considered a model selection criterion, AIC, to estimate the period for multiple periodic sequences. Similarly, for finite samples, a simulation study shows an improvement in period estimation from using information in a correlated sequence. The asymptotic properties of the AIC method will need further exploration.

Motivated by visualizing periodic sequences, the functional boxplot has been proposed as an informative exploratory tool for visualizing functional data, as well as its generalization, the enhanced functional boxplot. These functional boxplots were applied to sea surface temperatures, children growth and spatio-temporal precipitation datasets. With this new technique, outliers can be detected based on the 1.5 times the 50% central region empirical rule. Our approach is distinct from others in treating each curve as an observation rather than summarizing datasets on a pointwise basis. The descriptive statistics in a functional boxplot are rank-based, hence they may lead to building robust statistical models to capture the features of complex datasets. For spatio-temporal data, we have viewed the information as a temporal curve at each spatial location. An alternative would be to treat the dataset as a spatial surface at each time. In that case, we have proposed the surface boxplot which is a natural extension of the functional boxplot to $\mathbb{R}^3$.

We have also proposed a bootstrap method to adjust functional boxplots for correlations when visualizing functional and spatio-temporal data, as well as detecting

outliers. In a functional boxplot, potential outliers can be detected by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. However, for functional data, there is necessarily dependence in time for each curve. And for spatio-temporal data, curves from different locations are spatially correlated as well. Our simulation studies showed that the outlier detection performance is obviously affected by the dependence in time and space. Therefore, to correct the outlier detection performance, the constant factor of the empirical outlier rule is important. The factor $F = 1.5$ in a classical boxplot can be justified by a standard normal distribution, since it leads to a probability of 0.993 that any given observation is not an outlier. Following this idea, we proposed a simulation-based method to select this constant factor for a functional boxplot by controlling the probability of detecting no outliers to be 99.3% when actually no outliers are present. Then how to estimate the covariance function, especially for spatio-temporal data, is also important and robust techniques are needed when considering the potential presence of outliers in the original data. Furthermore, for spatio-temporal data, the constant factor can be also adjusted by the simulation-based method in the surface functional boxplot in $\mathbb{R}^3$ which leads to an adjusted surface functional boxplot.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium Information Theory*. Tsahkadsor, Armenian SSR, pp. 267–281.

Ammann, C. M., Washington, W. M., Meehl, G. A., Buja, L. & Teng, H. (2010). Climate engineering through artificial enhancement of natural forcings: Magnitudes and implied consequences. *Journal of Geophysical Research* **115**, D22109, doi:10.1029/2009JD012878.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods*. 2nd ed. New York: Springer Verlag.

Brown, P. E., Diggle, P. J., Lord, M. E. & Young, P. C. (2001). Space-time calibration of radar-rainfall data. *Applied Statistics* **50**, 221–242.

Campbell, M. J. & Walker, A. M. (1977). A survey of statistical work on the mackenzie river series of annual canadian lynx trappings for the years 1821-1934 and a new analysis. *Journal of the Royal Statistical Society, Series A* **140**, 411–431.

Collins, W. D. (2006). The community climate system model version 3 (ccsm3). *Journal of Climate* **19**, 2122–2143, doi:10.1175/JCLI3761.1.

Cressie, N. & Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**, 1330–1340.

Cressie, N. & Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. New York: Wiley.

Cuevas, A., Febrero, M. & Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics and Data Analysis* **51**, 1063–1074.

Cuevas, A., Febrero, M. & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics* **22**, 481–496.

Dioses, T., Dávalos, R. & Zuzunaga, J. (2002). El Niño 1982-1983 and 1997-1998: effects on peruvian jack mackerel and peruvian chub mackerel. *Investigaciones Marinas* **30**, 185–187.

Dryden, I. L. & Mardia, K. V. (1998). *Statistical Shape Analysis*. Chichester, U.K.: John Wiley and Sons.

Febrero, M., Galeano, P. & González-Manteiga, W. (2007). Functional analysis of NOx levels: location and scale estimation and outlier detection. *Computational Statistics* **22**, 411–427.

Febrero, M., Galeano, P. & González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics* **19**, 331–345.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer Verlag.

Fitzenberger, B., Koenker, R. & Machado, J. A. F. (2002). *Economic Applications of Quantile Regression*. New York: Springer Verlag.

Fletcher, P., Lu, C., Pizer, S. & Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging* **23**, 995–1005.

Fraiman, R. & Meloche, J. (1999). Multivariate L-estimation. *Test* **8**, 255–317.

Fraiman, R. & Muniz, C. (2001). Trimmed means for functional data. *Test* **10**, 419–440.

Genton, M. G. (1998). Highly robust variogram estimation. *Mathematical Geology* **30**, 213–221.

Genton, M. G. & Hall, P. (2007). Statistical inference for evolving periodic functions. *Journal of the Royal Statistical Society, Series B* **140**, 643–657.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**, 590–600.

Gneiting, T., Genton, M. G. & Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. In *Statistics of Spatio-Temporal Systems*. Eds. B. Finkenstaedt, L. Held & V. Isham. Chapman & Hall/CRC Press, pp. 151–175.

Hall, P. (2008). Nonparametric methods for estimating periodic functions, with applications in astronomy. Porto, Portugal, pp. 3–18.

Hall, P. & Li, M. (2006). Using the periodogram to estimate period in nonparametric regression. *Biometrika* **93**, 411–424.

Hall, P., Reimann, J. & Rice, J. (2000). Nonparametric estimation of a periodic function. *Biometrika* **87**, 545–557.

Hall, P. & Yin, J. (2003). Nonparametric methods for deconvolving multiperiodic functions. *Journal of the Royal Statistical Society, Series B* **65**, 869–886.

Handcock, M. S. & Wallis, J. R. (1994). An approach to spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association* **92**, 368–390.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* **41**, 190–195.

Hart, J. D. & Wehrly, T. E. (1993). Consistency of cross-validation when the data are curves. *Stochastic Processes and their Applications* **45**, 351–361.

Hyndman, R. J. & Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics* **19**, 29–45.

Intergovernmental Panel on Climate Change (IPCC) (2000). Special report on emission scenarios. In *A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*. Eds. N. Nakicenovic & R. Swart. Cambridge, U. K.: Cambridge University Press, p. 612.

Kyriakidis, P. C. & Journel, A. G. (1999). Geostatistical space-time models: a review. *Mathematical Geology* **31**, 651–684.

Liu, R. (1990). On a notion of data depth based on random simplices. *Annals of Statistics* **18**, 405–414.

Liu, R. Y., Parelius, J. M. & Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Annals of Statistics* **27**, 783–858.

López-Pintado, S. & Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association* **104**, 718–734.

Ma, Y. & Genton, M. G. (2000). Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis* **21**, 663–684.

Ma, Y. & Genton, M. G. (2001). Highly robust estimation of dispersion matrices. *Journal of Multivariate Analysis* **78**, 11–36.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. In *Proceedings of National Academy of Science of India*, vol. 12. Allahabad, India, pp. 49–55.

Maronna, R. A. & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* **50**, 295–304.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters* **1**, 327–332.

Quinn, B. G. & Thomson, P. J. (1991). Estimating the frequency of a periodic function. *Biometrika* **78**, 65–74.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. R Foundation for Statistical Computing.

Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. New York: Springer Verlag.

Rousseeuw, P., Ruts, I. & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician* **53**, 382–387.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–881.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Aplications*, G. S. Maddala & C. R. Rao, eds., vol. B. Amsterdam: Elsevier, pp. 101–121.

Schabenberger, O. & Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's Information Criterion. *Biometrika* **63**, 117–126.

Shumway, R. & Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*. New York: Springer.

Singh, K. (1991). A notion of majority depth. Technical Report, Department of Statistics, Rutgers University.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics* **16**, 155–160.

Sun, Y. & Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics* **20**, 316–334.

Sun, Y., Hart, J. D. & Genton, M. G. (2011). Nonparametric estimation of a periodic sequence. Manuscript, Department of Statistics, Texas A&M University.

Torrence, C. & Webster, P. (1999). Interdecadal changes in the ENOS-Monsoon system. *Journal of Climate* **12**, 2679–2690.

Tukey, J. W. (1970). *Exploratory Data Analysis*. Boston, MA: Addison-Wesley, Limited Preliminary Edition.

Tukey, J. W. (1975). Mathematics and picturing data. In *Proceedings of the International Congress of Mathematicians*, R. D. James, ed., vol. **2**. Canadian Mathematical Society, pp. 523–531.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Boston, MA: Addison-Wesley.

Vardi, Y. & Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1423–1426.

Woodroofe, M. (1982). On model selection and the arc sine laws. *Annals of Statistics* **10**, 1182–1194.

APPENDIX A

PROOF OF THEOREM 1

We begin with a lemma that addresses the behavior of the deterministic component of the CV criterion.

Lemma 1 *Define* $\mathcal{S}_n = \{q = 1, 2, \ldots, M_n : q \text{ is not a multiple of } p\}$, *and for* $q = 1, 2, \ldots,$

$$C_q = \frac{1}{n} \sum_{i=1}^{q} \sum_{j=1}^{k_{q,i}} (\mu_{i+(j-1)q} - \bar{\mu}_{qi})^2,$$

where $\bar{\mu}_{qi} = \sum_{j=1}^{k_{q,i}} \mu_{i+(j-1)q}/k_{q,i}$, $i = 1, \ldots, q$. *Then there exists* $\xi > 0$ *such that*

$$\min_{q \in \mathcal{S}_n} C_q > \xi$$

for all n *sufficiently large.*

Proof. The proof consists of two main steps:

Step 1 Show that $\min_{q \in \{2,\ldots,p-1\}} C_q$ is bounded away from 0 for all n sufficiently large.

Step 2 Argue that the bound in Step 1 can be applied to C_q for $q > p$ as well.

Step 1 Let q be one of $2, 3, \ldots, p-1$. Then there is an $\ell \in \{1, 2, \ldots, q\}$ such that not all of $\mu_\ell, \mu_{\ell+q}, \mu_{\ell+2q}, \ldots$ are the same. If there was not such an ℓ, then $\{\mu_j\}$ would be periodic of period q, which contradicts the assumption that p is the smallest period. We have

$$\begin{aligned} C_q &\geq \frac{1}{2nk_{q,q}} \sum_{j=1}^{k_{q,q}} \sum_{k=1}^{k_{q,q}} (\mu_{\ell+(j-1)q} - \mu_{\ell+(k-1)q})^2 \\ &= \frac{1}{2nk_{q,q}} \sum_{r=1}^{p} \sum_{s=1}^{p} n_{qr} n_{qs} (\mu_r - \mu_s)^2, \end{aligned}$$

where n_{qr} is the number of times $\mu_{\ell+(j-1)q}$ equals μ_r for j between 1 and $k_{q,q}$. Since not all of $\mu_{\ell+(j-1)q}$ are the same, there exist r_1 and r_2 such that $\mu_{r_1} \neq \mu_{r_2}$ and $n_{qr_1} > 0$ and $n_{qr_2} > 0$.

Obviously,

$$C_q \geq \frac{1}{2nk_{q,q}} n_{qr_1} n_{qr_2} (\mu_{r_1} - \mu_{r_2})^2.$$

The proof of Step 1 is done if we can show that each of n_{r_1} and n_{r_2} is bounded below by Cn, where $C > 0$. Let r be any integer between 1 and p such that $\mu_{\ell+(j-1)q}$ equals μ_r for some (smallest) j. It follows that there is a nonnegative integer k such that $\ell + (j-1)q - kp = r$. Therefore, for $m = 1, 2, \ldots$, we have

$$\ell + [(j-1) + mp]q - (k + mq)p = r,$$

which implies that $\mu_{\ell+sq} = \mu_r$ at $s = (j-1), (j-1)+p, (j-1)+2p, \ldots$ and hence that $n_{qr} \geq (k_{q,q} - j + 1)/p$. The result to be proven in Step 1 follows immediately.

Step 2 Suppose that $\{M_n\}$ is a sequence of integers such that $M_n \to \infty$ with $M_n = o(n)$. Now let $q = Mp + j$, where $1 \leq M \leq M_n$ and $1 \leq j \leq p-1$. We have

$$\begin{aligned}
C_q &= \frac{1}{n} \sum_{i=1}^{Mp+j} \sum_{k=1}^{k_{q,i}} (\mu_{i+(k-1)q} - \bar{\mu}_{qi})^2 \\
&= \frac{1}{2n} \sum_{i=1}^{Mp+j} \frac{1}{k_{q,i}} \sum_{k=1}^{k_{q,i}} \sum_{\ell=1}^{k_{q,i}} (\mu_{i+(k-1)q} - \mu_{i+(\ell-1)q})^2 \\
&\geq \frac{1}{2nk_{q,q}} \sum_{i=1}^{Mp+j} \sum_{k=1}^{k_{q,q}} \sum_{\ell=1}^{k_{q,q}} (\mu_{i+(k-1)j} - \mu_{i+(\ell-1)j})^2 \\
&= \sum_{m=1}^{M} C_{qm} + \frac{1}{2nk_{q,q}} \sum_{i=Mp+1}^{Mp+j} \sum_{k=1}^{k_{q,q}} \sum_{\ell=1}^{k_{q,q}} (\mu_{i+(k-1)j} - \mu_{i+(\ell-1)j})^2,
\end{aligned}$$

where

$$C_{qm} = \frac{1}{2nk_{q,q}} \sum_{i=(m-1)p+1}^{mp} \sum_{k=1}^{k_{q,q}} \sum_{\ell=1}^{k_{q,q}} (\mu_{i+(k-1)j} - \mu_{i+(\ell-1)j})^2$$
$$= \frac{1}{2nk_{q,q}} \sum_{r=1}^{p} \sum_{k=1}^{k_{q,q}} \sum_{\ell=1}^{k_{q,q}} (\mu_{r+(k-1)j} - \mu_{r+(\ell-1)j})^2.$$

It follows that $C_q \geq MC_{q1}$, and hence that

$$C_q \geq \frac{M}{2nk_{q,q}} \sum_{r=1}^{j} \sum_{k=1}^{k_{q,q}} \sum_{\ell=1}^{k_{q,q}} (\mu_{r+(k-1)j} - \mu_{r+(\ell-1)j})^2.$$

For future reference we note that

$$k_{q,q} = \left[\frac{n}{q}\right] = \left[\frac{n}{Mp+j}\right] \geq \frac{n}{p(M_n+1)} - 1. \tag{A.1}$$

Using the same type of notation and arguing exactly as in the proof of Step 1,

$$C_q \geq \frac{M}{2nk_{q,q}} n_{q1} n_{q2} (\mu_{r_1(j)} - \mu_{r_2(j)})^2$$
$$\geq \frac{M}{2nk_{q,q}} n_{q1} n_{q2} \delta,$$

where δ is the smallest nonzero value of $(\mu_i - \mu_j)^2$ for i, j in $\{1, \ldots, p\}$. Now, for $i = 1, 2$, n_{qi} is the number of times that $\mu_{r+(k-1)j}$ equals $\mu_{r_i(j)}$ as k ranges between 1 and $k_{q,q}$. As in the proof of Step 1, we know that

$$n_{qi} \geq \left(\frac{k_{q,q} - k(i,j) + 1}{p}\right),$$

where $k(i,j)$ is the smallest k for which $\mu_{r+(k-1)j} = \mu_{r_i(j)}$. Importantly, $k(i,j)$ depends on j but not M. We thus have

$$C_q \geq \frac{\delta M}{2np^2 k_{q,q}} (k_{q,q} - k^* + 1)^2$$
$$\geq \frac{\delta M}{2(Mp+j)p^2} \left(1 - \frac{Mp+j}{n}\right) \left(1 - \frac{k^*-1}{k_{q,q}}\right)^2,$$

where k^* is the largest of the integers $k(i,j)$, $i = 1, 2$, $j = 1, \ldots, p-1$.

Now,

$$\frac{M}{Mp+j} \geq \frac{M}{p(M+1)} \geq \frac{1}{2p},$$

and hence

$$C_q \geq \frac{\delta}{4p^3}\left(1 - \frac{Mp+j}{n}\right)\left(1 - \frac{k^*-1}{k_{q,q}}\right)^2.$$

Recalling (A.1) and the fact that $M_n = o(n)$, it follows that $(Mp+j)/n$ and $(k^*-1)/k_{q,q}$ are smaller than $1/2$ for all n sufficiently large, and so

$$C_q \geq \frac{\delta}{32p^3}$$

for all n sufficiently large. □

To prove that $\lim_{n\to\infty} P(\hat{p} \in \mathcal{S}_n) = 0$ we must show that

$$\lim_{n\to\infty} P\left[\bigcap_{1\leq\ell\leq M_n/p}\ \bigcap_{q\in\mathcal{S}_n} \{\mathrm{CV}(q) - \mathrm{CV}(\ell p) > 0\}\right] = 1.$$

The periodicity of $\boldsymbol{\mu}$ entails that

$$\mathrm{CV}(q) - \mathrm{CV}(\ell p) = \frac{1}{n}(S_{q,n} - S_{\ell p,n}) + \frac{2}{n}\sum_{i=1}^{q}\sum_{j=1}^{k_{q,i}}(\varepsilon_{qij} - \bar{\varepsilon}^j_{qi})(\mu_{qij} - \bar{\mu}^j_{qi}) + \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{k_{q,i}}(\mu_{qij} - \bar{\mu}^j_{qi})^2,$$

where, for each q, $S_{q,n} = \sum_{i=1}^{q}\sum_{j=1}^{k_{q,i}}(\varepsilon_{qij} - \bar{\varepsilon}^j_{qi})^2$. Using the result of Lemma 1, it is easily checked that, for all n sufficiently large,

$$\min_{q\in\mathcal{S}_n} \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{k_{q,i}}(\mu_{qij} - \bar{\mu}^j_{qi})^2 \geq \frac{\xi}{2}.$$

It therefore follows that

$$\begin{aligned} P\left[\bigcap_{1\le\ell\le M_n/p}\bigcap_{q\in\mathcal{S}_n}\{\mathrm{CV}(q)-\mathrm{CV}(\ell p)>0\}\right] &\ge 1-\sum_{\ell}\sum_{q}P\left(\frac{1}{n}(S_{q,n}-S_{\ell p,n})\le-\frac{\xi}{6}\right) \\ &- \sum_{\ell}\sum_{q}P\left(A_{q,n}\le-\frac{\xi}{6}\right) \\ &- \sum_{\ell}\sum_{q}P\left(B_{q,n}\le-\frac{\xi}{6}\right), \end{aligned} \tag{A.2}$$

where

$$A_{q,n}=\frac{2}{n}\sum_{i=1}^{q}\sum_{j=1}^{k_{q,i}}\varepsilon_{qij}(\mu_{qij}-\bar{\mu}_{qi}^{j}) \quad\text{and}\quad B_{q,n}=-\frac{2}{n}\sum_{i=1}^{q}\sum_{j=1}^{k_{q,i}}\bar{\varepsilon}_{qi}^{j}(\mu_{qij}-\bar{\mu}_{qi}^{j}).$$

For t any positive number, Bernstein's form of the Markov inequality implies that

$$P\left(A_{q,n}\le-\frac{\xi}{6}\right)\le\exp\left[-nt\left(\frac{\xi}{6}-8t\sigma^2B\right)\right],$$

where $B=\max_{1\le i\le p}\mu_i^2$. Taking t to be smaller than $\xi/(48B\sigma^2)$, it is now clear that there exists a positive number C such that $P(A_{q,n}\le-\xi/6)\le e^{-Cn}$, and since M_n is smaller than n, $\sum_\ell\sum_q P(A_{q,n}\le-\xi/6)\to 0$ as $n\to\infty$. The other two terms on the right hand side of (A.2) can be dealt with in the same way, and therefore $\lim_{n\to\infty}P(\hat{p}\in\mathcal{S}_n)=0$.

Now we consider $P(\hat{p}=p)$, which, by the result just proven, is asymptotically equal to

$$P\left(\bigcap_{2\le m\le M_n/p}\{\mathrm{CV}(mp)-\mathrm{CV}(p)>0\}\right). \tag{A.3}$$

(The proof for other $P(\hat{p}=jp)$ is similar and hence omitted.) We have

$$\mathrm{CV}(mp)-\mathrm{CV}(p)=\frac{1}{n}\sum_{i=1}^{mp}C_{mp,i}\left\{\sum_{j=1}^{k_{mp,i}}\varepsilon_{mp,ij}^2-k_{mp,i}\bar{\varepsilon}_{mp,i}^2\right\}-\frac{1}{n}\sum_{i=1}^{p}C_{p,i}\left\{\sum_{j=1}^{k_{pi}}\varepsilon_{pij}^2-k_{pi}\bar{\varepsilon}_{pi}^2\right\},$$

where $C_{q,i} = k_{q,i}^2/(k_{q,i}-1)^2$. Writing $\hat{\sigma}^2 = \sum_{i=1}^n \varepsilon_i^2/n$, we have

$$n(\mathrm{CV}(mp) - \mathrm{CV}(p)) = \sigma^2 \sum_{i=1}^{p} C_{p,i} U_{1,i}^2 - \sigma^2 \sum_{i=1}^{mp} C_{mp,i} U_{m,i}^2 + n\sigma^2 (C_{mp,1} - C_{p,1}) + R_{m,n},$$

where $U_{m,i} = \sqrt{k_{mp,i}}\bar{\varepsilon}_{mp,i}/\sigma$, $i = 1, \ldots, mp$, $m = 1, \ldots, M_n$, and

$$R_{m,n} = n(\hat{\sigma}^2 - \sigma^2)(C_{mp,1} - C_{p,1}) + \sum_{i=1}^{mp} (C_{mp,i} - C_{mp,1}) \sum_{j=1}^{k_{mp,i}} \varepsilon_{mp,ij}^2 + \sum_{i=1}^{mp} (C_{p,1} - C_{p,i}) \sum_{j=1}^{k_{p,i}} \varepsilon_{p,ij}^2.$$

We remark that, for each m, $U_{m,1}^2, \ldots, U_{m,mp}^2$ are i.i.d. χ_1^2 random variables. Subsequently we use the following facts:

$$n(C_{mp,1} - C_{p,1}) = 2p(m-1) + O\left(\frac{M_n^2}{n}\right), \quad C_{mp,1} = 1 + O\left(\frac{M_n}{n}\right), \quad C_{p,1} = 1 + O\left(\frac{1}{n}\right),$$

$$|C_{mp,i} - C_{mp,1}| \le O\left(\frac{M_n}{n}\right)^2 \quad \text{and} \quad |C_{p,i} - C_{p,1}| \le O\left(\frac{1}{n}\right)^2,$$

with the last two inequalities holding uniformly in i, since, for each q, $k_{q,1}, \ldots, k_{q,q}$ take on at most two distinct values that differ by only 1.

So now we have

$$\frac{n(\mathrm{CV}(mp) - \mathrm{CV}(p))}{\sigma^2} = \sum_{i=1}^{p} U_{1,i}^2 - \sum_{i=1}^{mp} U_{m,i}^2 + 2p(m-1) + \tilde{R}_{m,n},$$

where, using previously stated facts, we have

$$\tilde{R}_{m,n} = \frac{R_{m,n}}{\sigma^2} + O_p\left(\frac{M_n^2}{n}\right).$$

Note that the term $|R_{m,n}|$ is bounded almost surely by

$$n(C_{mp,1} - C_{p,1})|\hat{\sigma}^2 - \sigma^2| + n\hat{\sigma}^2 \left[\max_i |C_{mp,i} - C_{mp,1}| + \max_i |C_{p,1} - C_{p,i}|\right],$$

which is $O_p(M_n/\sqrt{n}) + O_p(M_n^2/n)$.

Now let $\{\delta_n\}$ be an arbitrary sequence of positive numbers that tend to 0. The sequence of probabilities (A.3) may be bounded above and below by probability se-

quences that have the same limits as

$$P\left(\bigcap_{2\leq m\leq M_n/p}\left\{\sum_{i=1}^{p}U_{1,i}^2-\sum_{i=1}^{mp}U_{m,i}^2+2p(m-1)>-\delta_n\right\}\right)$$

and

$$P\left(\bigcap_{2\leq m\leq M_n/p}\left\{\sum_{i=1}^{p}U_{1,i}^2-\sum_{i=1}^{mp}U_{m,i}^2+2p(m-1)>\delta_n\right\}\right),$$

respectively. The last statement is proven by applying Bernstein's form of the Markov inequality to the sequences $P(\bigcap_{2\leq m\leq M_n/p}\tilde{R}_{m,n}>\delta_n)$ and $P(\bigcap_{2\leq m\leq M_n/p}\tilde{R}_{m,n}>-\delta_n)$, and using the assumption that $M_n=o(\sqrt{n})$ and the fact that δ_n can be defined to converge arbitarily slowly to 0. It is now clear that

$$\lim_{n\to\infty}P(\hat{p}=p)=\lim_{n\to\infty}P\left(\bigcap_{2\leq m<M_n/p}\left\{\sum_{i=1}^{p}U_{1,i}^2-\sum_{i=1}^{mp}U_{m,i}^2+2p(m-1)>0\right\}\right).$$

The result will be proven if we can verify that the $U_{m,i}$s have the same limiting correlation structure as that of the $Z_{m,i}$s. By construction, $U_{m,1},\ldots,U_{m,mp}$ are mutually independent. Now let $r>m$ and consider

$$\begin{aligned}\mathrm{Corr}(U_{m,i},U_{r,j}) &= \frac{1}{\sqrt{k_{mp,i}k_{rp,j}}\sigma^2}\sum_{\ell=1}^{k_{mp,i}}\sum_{s=1}^{k_{rp,j}}E\left[\varepsilon_{i+mp(\ell-1)}\varepsilon_{j+rp(s-1)}\right]\\ &= \frac{1}{\sqrt{k_{mp,i}k_{rp,j}}}N_{m,r,i,j},\end{aligned}$$

where $N_{m,r,i,j}$ is the number of times that $j+rp(s-1)=i+mp(\ell-1)$. Now, if there is a pair (ℓ,s) that satisfies this equation, then $(\ell+\eta r,s+\eta m)$, $\eta=1,2,\ldots$ are also

solutions and $N_{m,r} \sim k_{rp,j}/m$. Therefore, when $N_{m,r} > 0$,

$$\begin{aligned}
\mathrm{Corr}(U_{m,i}, U_{r,j}) &\sim \frac{k_{rp,j}}{m\sqrt{k_{mp,i}k_{rp,j}}} \\
&= \frac{1}{m}\sqrt{\frac{k_{rp,j}}{k_{mp,i}}} \\
&\sim \frac{1}{m}\sqrt{\frac{n/(rp)}{n/(mp)}} \\
&= \frac{1}{\sqrt{mr}}.
\end{aligned}$$

□

APPENDIX B

PROOF OF MLES OF SEQUENCE MEANS

For the bivariate case, suppose we observe bivariate sequences from the following model:

$$\begin{cases} X_t = \mu_{1t} + \varepsilon_{1t}, \\ Z_t = \mu_{2t} + \varepsilon_{2t}, \end{cases}$$

where $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}_t \sim i.i.d.N(\mathbf{0}, \Sigma)$, and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

The likelihood is of the form

$$L = \frac{1}{1-\rho^2}\left\{ \frac{1}{\sigma_1^2}\sum_{t=1}^{n}(X_t - \mu_{1,t})^2 + \frac{1}{\sigma_2^2}\sum_{t=1}^{n}(Z_t - \mu_{2,t})^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{t=1}^{n}(X_t - \mu_{1,t})(Z_t - \mu_{2,t}) \right\}. \tag{B.1}$$

Let q_1 and q_2 be the period candidates for X_t and Z_t respectively and $m_i = \mu_{2,i+q_2} = \mu_{2,i+2q_2} = \cdots$ for $i = 1, \ldots, q_2$, $\eta_i = \mu_{1,i+q_1} = \mu_{1,i+2q_1} = \cdots$ for $i = 1, \ldots, q_1$. First, minimize

$$\frac{1}{\sigma_2^2}\sum_{t=1}^{n}(Z_t - \mu_{2,t})^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{t=1}^{n}(X_t - \mu_{1,t})(Z_t - \mu_{2,t}) \tag{B.2}$$

with respect to $m_1, \ldots, m_{q_2}$, where $m_i = \mu_{2,i+q_2} = \mu_{2,i+2q_2} = \cdots$ for $i = 1, \ldots, q_2$.

Define $Z_{q_2ij} = Z_{i+(j-1)q_2}$, where $i = 1, \ldots, q_2$ and k_{q_2i} is the largest integer such that $i + q_2 k_{q_2,i} \leq n$. Then (B.2) can be written as

$$A = \frac{1}{\sigma_2^2}\sum_{i=1}^{q_2}\sum_{j=1}^{k_{q_2i}}(Z_{q_2ij} - m_i)^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^{q_2}\sum_{j=1}^{k_{q_2i}}(X_{q_2ij} - \mu_{1,q_2ij})(Z_{q_2ij} - m_i). \tag{B.3}$$

The first derivative

$$\frac{\partial A}{\partial m_i} = -\frac{2}{\sigma_2^2}\sum_{j=1}^{n_{k_2i}}(Z_{q_2ij} - m_i) + \frac{2\rho}{\sigma_1\sigma_2}\sum_{j=1}^{k_{q_2i}}(X_{q_2ij} - \mu_{1,q_2ij}) = 0$$

gives

$$m_i = \bar{Z}_{q_2 i} - \rho\frac{\sigma_2}{\sigma_1}(\bar{X}_{q_2 j} - \bar{\mu}_{1,q_2 i}), \tag{B.4}$$

where $\bar{Z}_{q_2 i}$ is the average of $Z_{q_2 i\ell}$, $\ell = 1, \ldots, k_{q_2 i}$.

Then plug (B.4) into equation (B.1),

$$\begin{aligned} L_1 =& \frac{1}{\sigma_1^2}\sum_{i=1}^{q_1}\sum_{j=1}^{k_{q_1 i}}(X_{q_1 ij} - \eta_i)^2 + \frac{1}{\sigma_2^2}\sum_{i=1}^{q_2}\sum_{j=1}^{k_{q_2 i}}(Z_{q_2 ij} - \bar{Z}_{q_2 i})^2 \\ & - \frac{2\rho^2}{\sigma_1}\sum_{i=1}^{q_2} k_{q_2 i}(\bar{X}_{q_2 i} - \bar{\mu}_{q_2 i})^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum_{i=1}^{q_2}\sum_{j=1}^{k_{q_2 i}}(X_{q_2 ij} - \mu_{1,q_2 ij})(Z_{q_2 ij} - \bar{Y}_{q_2 i}). \end{aligned}$$

We then have

$$\partial\mu_{1,q_2 ij}/\partial\eta_r = 1, \text{ if } i + (j-1)q_2 = r + (\ell - 1)q_1 \text{ for some } \ell, \text{ otherwise } 0,$$

and $\partial\bar{\mu}_{q_2 i}/\partial\eta_r = N_{ir}/k_{q_2 i}$, where N_{ir} is the number of times $i+(j-1)q_2 = r+(\ell-1)q_1$ when j ranges from 1 to $k_{q_2 i}$ and ℓ ranges from 1 to $k_{q_1 r}$.

Then,

$$\begin{aligned} \frac{\partial L_1}{\partial\eta_r} =& -\frac{2}{\sigma_1^2}k_{q_1 r}(\bar{X}_{q_1 r} - \eta_r) \\ & + \frac{2\rho^2}{\sigma_1^2}\left(\sum_{i=1}^{q_1}\bar{X}_{q_1 i}N_{ir} - \sum_{i=1}^{q_2}\frac{1}{k_{q_2 i}}\sum_{k=1}^{q_1}N_{ir}N_{ik}\eta_k\right) \\ & + \frac{2\rho}{\sigma_1\sigma_2}S_{q_1,q_2,r}, \end{aligned}$$

where $S_{q_1,q_2,r} = \sum_{i=1}^{k_{q_1 r}}\left(Y_{q_1 rj} - \bar{Y}_{q_1 rj}^{q_2}\right)$.

We can see that $\partial L_1/\partial\eta_r = 0$ if and only if

$$k_{q_1 r}(\bar{X}_{q_1 r} - \eta_r) - \rho^2\frac{1}{k_{q_2 i}}\sum_{k=1}^{q_1}N_{ir}N_{ik}\eta_k = -\rho^2\sum_{i=1}^{q_1}\bar{X}_{q_1 i}N_{ir} - \rho\frac{\sigma_1}{\sigma_2}S_{q_1,q_2,r}.$$

This yields a set of q_1 linear equations. Arrange them so that we can solve a smaller

system, i.e., $q_1 < q_2$. The coefficient of η_r in the r-th equation is

$$k_{q_1 r} - \rho^2 \sum_{i=1}^{q_2} \frac{1}{k_{q_2 i} N_{ir}^2},$$

and the coefficient of η_k $(k \neq r)$ is

$$-\rho^2 \sum_{i=1}^{q_2} \frac{1}{k_{q_2 i}} N_{ir} N_{ik}.$$

The equations are

$$B_{q_1,q_2} \boldsymbol{\eta} = \mathbf{b}_{q_1,q_2},$$

where $B_{q_1,q_2}(j,k) = \rho^2 \sum_{i=1}^{q_2} \frac{1}{k_{q_2 i}} N_{ij} N_{ik} - k_{q_1 j} I(j-k)$ for $I(0) = 1$, 0 otherwise, and

$$\mathbf{b}_{q_1,q_2}(j) = \rho^2 \sum_{i=1}^{q_2} \bar{X}_{q_2 i} N_{ij} + \rho \frac{\sigma_1}{\sigma_2} S_{q_1,q_2,j} - k_{q_1,j} \bar{X}_{q_1 j}.$$

When $q_1 = q_2$, the MLEs of $\eta_1, \ldots, \eta_{q_1}$ and $m_1, \ldots, m_{q_2}$ are the usual stacked means.

□

VITA

Name	Ying Sun
Research Interests	Spatial and spatio-temporal statistics with environmental and climate applications, nonparametric function estimation, functional data analysis, time series, multivariate analysis and data mining.
Education	Ph.D., Statistics, Aug 2011, Texas A&M University, College Station, TX. Advisors: Jeffrey D. Hart & Marc G. Genton M.S., Statistics, Jul 2006, Tsinghua University, Beijing, China
Address	Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX, 77843
Email	sunwards@stat.tamu.edu
Webpage	http://www.stat.tamu.edu/∼sunwards
Organizations	American Statistical Association member (ASA) Institute of Mathematical Statistics member (IMS)