ESTIMATING LYAPUNOV EXPONENTS IN CHAOTIC TIME SERIES WITH LOCALLY WEIGHTED REGRESSION

by

Zhan-Qian Lu

(dissertation)

ESTIMATING LYAPUNOV EXPONENTS IN CHAOTIC TIME SERIES WITH LOCALLY WEIGHTED REGRESSION

by

Zhan-Qian Lu

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics.

Chapel Hill

1994

Approved by:

______________________________ Advisor

______________________________ Reader

______________________________ Reader

Zhan-Qian Lu. Estimating Lyapunov Exponents in Chaotic Time Series with Locally Weighted Regression. (Under the direction of Richard L. Smith.)

ABSTRACT

Nonlinear dynamical systems often exhibit chaos, which is characterized by sensitive dependence on initial values or more precisely by a positive Lyapunov exponent. Recognizing and quantifying chaos in time series represents an important step toward understanding the nature of random behavior and revealing the extent to which short-term forecasts may be improved. We will focus on the statistical problem of quantifying chaos and nonlinearity via Lyapunov exponents.

Predicting the future or determining Lyapunov exponents requires estimation of an autoregressive function or its partial derivatives from time series. The multivariate locally weighted polynomial fit is studied for this purpose. In the nonparametric regression context, explicit asymptotic expansions for the conditional bias and conditional covariance matrix of the regression and partial derivative estimators are derived for both the local linear fit and the local quadratic fit. These results are then generalized to the time series context. The joint asymptotic normality of the estimators is established under general short-range dependence conditions, where the asymptotic bias and asymptotic covariance matrix are explicitly given. We also discuss extension to fractal time series, where the finite-dimensional probability measure is only assumed to possess a fractal dimension and can be singular with respect to the Lebesgue measure.

The results on partial derivative estimation are subsequently applied to estimation of Lyapunov exponents. Using the asymptotic theory of the eigenvalues from a random matrix, we are able to characterize the asymptotic distribution for the estimators of local Lyapunov exponents. The local Lyapunov exponents may provide a way of quantifying short-term predictability in a system. The results may shed some light on the estimators of Lyapunov exponents.

ACKNOWLEDGMENTS

I wish to thank my advisor, Professor Richard L. Smith, first of all for showing me how to do research, for his generous advices and kind encouragements at all times. It is also very fortunate to have the other members of my advisory committee Professors Jianqing Fan, Gopinath Kallianpur, Steve Marron, and Doug Nychka, whose insightful comments and valuable suggestions are gratefully appreciated.

I have enjoyed stimulating discussions with Professors Steve Ellner, Doug Nychka, Ron Gallant and their students. I thank Prof. Steve Ellner for his interest and encouragement. I am particularly grateful to Prof. Doug Nychka for his valuable advices and constant support.

I have useful discussions with Professors Steve Marron and Jianqing Fan, whose influence in my understanding of nonparametric regression can be seen throughout this work. I thank Professor Chuanshu Ji for his earlier interest in this work.

It is a pleasure to thank the terrific staff in this department for making the day-to-day matters appear so easy. I would also like to thank the department for providing a wonderful environment for study and financial support.

Contents

1 INTRODUCTION **1**

1.1 Introduction 1

1.2 What is Chaos? 5

1.3 Chaotic Time Series 8

1.4 Nonparametric Regression 10

1.5 Estimation of Lyapunov Exponents 11

1.6 Organization of Chapters 12

2 THEORY OF CHAOS **14**

2.1 Introduction 14

2.2 Some Chaos Examples 16

2.3 Strange Attractor 22

2.4 Lyapunov Exponents 27

2.5 Noisy Systems 30

2.6 Some Advanced Theory 32

2.7 Fractal Dimension 36

2.8 Entropy 40

2.9 State Space Reconstruction 41

3 MULTIVARIATE LOCALLY WEIGHTED REGRESSION 47

3.1 Introduction 47

3.2 Some Notations 51

3.3 Local Linear Fit 53

3.4 Local Quadratic Fit 57

3.5 Discussion 62

3.6 Proofs 64

4 NONPARAMETRIC ESTIMATION WITH TIME SERIES 79

4.1 Introduction 79

4.2 Regression-type Time Series Model 81

4.3 Local Linear Fit 84

4.4 Local Quadratic Fit 85

4.5 Nonparametric Estimation with Fractal Time Series 86

4.6 Deterministic Fit 94

4.7 Proofs 95

5 ESTIMATION OF LYAPUNOV EXPONENTS 105

5.1 Introduction 105

5.2 Estimation of Jacobian Matrix 107

5.3 Asymptotic Distribution of Eigenvalues 111

5.4 Asymptotic Theory of Singular Values 119

5.5 Application to Estimation of Singular Values of Jacobian Product. . . 121

5.6 Estimation of Local Lyapunov Exponents 123

6 OUTLOOK 127

Chapter 1

INTRODUCTION

1.1 Introduction

A time evolution is often described by the solution to difference (discrete-time) or differential (continuous-time) equations. A nonlinear dynamical system can generate orbits which are typically aperiodic, irregular, and apparently random. This phenomenon is popularly called *chaos*. Chaos has provided an alternative interpretation of the erratic or random behavior in natural processes. Chaos theory has been applied to many different fields including physics (e.g. in the study of turbulence, Eckmann and Ruelle 1985), meteorology (Lorenz 1963), ecology, biology, epidemiology (e.g. Sugihara and May 1990), economics, and finance. Recently, much interest has focused on statistical analysis of chaotic time series. Chaotic time series have potentially wide applications. Our main concern in this study is to provide the necessary statistical theory for quantifying chaos, with focus on the Lyapunov exponent approach. Since typically only limited noisy observations are available, the statistical study of available chaos methods are crucial to their proper applications (e.g. Ruelle 1990, Smith 1992b).

A real system may be subject to various sources of small random noises, such as computer roundoff error, measurement error, external environmental noise, etc. So a reasonable model may be given by a noisy system, or a deterministic system subject to small dynamical noises. See e.g. Smith (1992b), Nychka et al (1992).

The connection of chaos and dynamical systems to traditional nonlinear time series analysis is discussed by Tong (1990). Most ideas in deterministic chaos can be extended to analyzing a noisy system. Consideration of chaos in a stochastic system enables us to focus on the deterministic structure, particularly when it is the dominating mechanism of the dynamical behavior. Thus, once chaos is found to be present in a system, much more structure is known about the random dynamical behavior, which will be particularly useful for short-term prediction. Chaotic time series may be seen as consisting of three interrelated parts, detecting chaos, modeling, and nonlinear prediction.

Recognizing and quantifying chaos in time series represents an important step toward understanding the nature of random behavior and revealing the extent to which short-term forecasts may be improved. Several approaches have been proposed, including estimating fractal dimensions (e.g. Smith 1991, 1992a), nonlinear forecasting (Sugihara and May 1990, and Casdagli 1992), estimating entropy (e.g. Eckmann and Ruelle 1985), and estimating Lyapunov exponents (Wolf et al 1985, Abarbanel et al 1991, McCaffrey et al 1992).

Among the methods proposed, dimension estimation is perhaps the simplest approach. It provides a test about the finite dimensionality of a system. However, the dimension estimates may be sensitive to substantial amount of measurement errors in data, and may get even worse with the dynamical noises considered here, see e.g. Smith (1992a). Similar difficulty may be expected for entropy estimates (e.g. those in Eckmann and Ruelle 1985).

With the approach of determining Lyapunov exponents or assessing nonlinear forecasts, this problem can be avoided. Chaos is defined by the sensitive dependence on initial values or the existence of a positive Lyapunov exponent. Consequently, long-term prediction is impossible in a chaotic system since any uncertainty at the initial step will be amplified exponentially fast. However, the presence of chaos also reveals the possibility that short-term forecasts may be improved through nonlinear methods. This in turn gives a test of nonlinearity through evaluating the quality of nonlinear forecasts.

In addition, the Lyapunov spectrum is closely related to fractal dimension and entropy. So we will focus on the approach of estimating Lyapunov exponents. For detecting chaos, estimating the dominant Lyapunov exponent is most important. We also consider estimation of other Lyapunov exponents.

The Jacobian-based approach should be used in estimating Lyapunov exponents in a noisy system, see e.g. McCaffrey (1992). The estimators of Lyapunov exponents are given in terms of the growth rates of the singular values of the estimated l-step Jacobian matrix product of the system. For consistency of estimating Lyapunov exponents, l should be chosen to depend on sample size n and tend to infinity as n does. There are difficulties in establishing the convergence rate, which is also tied with the choice of l, as discussed in Ellner et al (1991), McCaffrey et al (1992).

To provide some insights on the problem, we investigate the following simpler problem: what is the asymptotic behavior of the estimators for any fixed l? We expect that the derived results will shed some light on the estimators of Lyapunov exponents. This problem maybe worth studying in itself for its relevance to characterizing predictability in time series. Since chaotic time series are well-known for their long-term unpredictability, considerable interest has focused on short-term predictability, e.g. this is of much interest in weather forecasting. The *local Lyapunov exponents*, which are defined as the finite-time average divergence rates, may provide more relevant measures of predictability. A similar idea has also been used in Wolff (1992), who has considered the one-dimensional case, and Bailey (1993), who has focused on the maximum local Lyapunov exponent.

Nonlinear prediction or determining the Lyapunov exponents requires estimation of the nonlinear autoregressive function or its partial derivatives. We will focus on the nonparametric method of *locally weighted polynomial fit.* Most recent papers in chaos literature, such as Eckmann and Ruelle (1985), Farmer and Sidorowich (1987), Sugihara and May (1990), Casdagli (1992), and Abarbanel et al (1991), have considered some versions of nonparametric techniques, particularly the *local polynomial method.* The local polynomial method, or the locally weighted polynomial fit in its modern form, is also of current interest in statistics literature, such as in Fan (1993) and Ruppert and Wand (1994). We will study the multivariate locally weighted polynomial fit, with emphasis on partial derivative estimation.

There is a large literature on the mathematical aspects of chaos and dynamical systems, including Guckenheimer and Holmes (1990), Eckmann and Ruelle (1985), Devaney (1989), Sinai (1989), Ruelle (1989), and Ott (1993). Numerical implementation is discussed in Parker and Chua (1989). Eckmann and Ruelle (1985) also give a good review on some past work on the statistical aspects of chaos. Much statistical work has appeared in physics literature such as Physica D. It is relatively recent that chaos has attracted attention from the statistical community. See the special issue on chaos of JRSS, Series B, 1992, which includes Nychka et al (1992), Smith (1992a), Casdagli (1992), and Wolff (1992). See also a discussion paper by Bartlett (1990), two reviews by Berliner (1992), Chatterjee and Yilmaz (1992) and accompanying discussions therein.

This chapter is organized as follows. An introduction to the basic concepts of chaos and dynamical systems is given in Section 1.2, where the Lyapunov exponents are defined. Section 1.3 discusses the problem of detecting chaos in time series, with focus on the approach of estimating Lyapunov exponents. The statistical problem of estimating an autoregressive function or its partial derivatives is discussed in the nonparametric regression setup, which is introduced in Section 1.4. The locally weighted polynomial fit is emphasized. The application of partial derivative estimation to estimating Lyapunov exponents is discussed in Section 1.5, where we define a finite-time version of the (global) Lyapunov exponents, which we call the local Lyapunov exponents. Finally, the organization of other chapters is given in Section 1.6.

1.2 What is Chaos?

A time evolution is often described by a dynamical system, which is given by the solution to ordinary differential equations (ODE) in continuous time

$$d\mathbf{x}(t)/dt = M(\mathbf{x}(t)), \text{ where } \mathbf{x}(t) \in \mathcal{R}^p \text{ for } t \in [0, \infty), \tag{1.1}$$

or difference equations in discrete time

$$\mathbf{x}_{n+1} = M(\mathbf{x}_n), \text{ where } \mathbf{x}_n \in \mathcal{R}^p \text{ for integers } n, \tag{1.2}$$

where M is a $\mathcal{R}^p \rightarrow \mathcal{R}^p$ map. It is called a differentiable dynamical system if M is differentiable. It is emphasized that the above system is finite-dimensional. Later on, we will consider a noisy system, which is, strictly speaking, infinite dimensional. However, even in that case, our focus is still on the detection of a finite-dimensional structure, and the infinite-dimensional noise component plays a minor role.

A nonlinear deterministic system can produce typically aperiodic, irregular, and apparently random orbit, a phenomenon often called chaos. For example, the system $x_{n+1} = \mu x_n(1 - x_n)$ is known to go through a whole spectrum of simple (fixed or periodic point) and complex dynamics (chaos) as μ varies over $[0, 4]$ (Devaney, 1989, Sinai 1989). A simple higher-dimensional example is given by the Hénon map $M(x, y) = (by + 1 - ax^2, x)$, where a, b are constants. Numerical study shows that the Hénon map has complicated dynamics for parameter values $a = 1.4, b = .3$. More examples are given in Section 2.2. We will be mainly interested in dissipative systems, for which a main concept is the *attractor*, a limiting set A on which a trajectory eventually settles down for a typical initial value. An *asymptotic measure* can also be defined on an attractor. See Section 2.3 for more details.

A chaotic system is characterized by the exponentially fast divergence of nearby trajectories for most initial values, or the prevalence of *sensitive dependence on initial conditions*. This notion is quantified by *Lyapunov exponents*, which are defined as the average exponential rates of divergence or convergence of nearby trajectories. Visually, imagine we can monitor and plot the evolution of a small sphere in the phase space. The sphere will become an ellipsoid due to the local stretching and contracting nature of the system, and the major axis of the ellipsoids will rapidly (at an exponential speed) approach the size of the attractor. So at each point, corresponding to different directions in the phase space, there may exist different divergence or convergence rates, implying that there are p Lyapunov exponents in total, which we denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The ith Lyapunov exponent λ_i is defined by the average exponential rate at which the ith principal axis of the ellipsoids expands or contracts for any i between 1 and p.

For example, if $\mathbf{x}$ is a fixed point ($M\mathbf{x} = \mathbf{x}$) or periodic point with period n (i.e. $M^n\mathbf{x} = \mathbf{x}$) of a map M, the Lyapunov exponents are given by the logarithm of the modulus of the eigenvalues of the Jacobian matrix $D_M(\mathbf{x})$ or $D_{M^n}(\mathbf{x})$. Now we will define the Lyapunov exponents for an aperiodic orbit.

Consider the discrete-time system (1.2). Let $T(\mathbf{x}) = D_M(\mathbf{x})$ denote the Jacobian matrix of M at $\mathbf{x}$. The idea is to study the divergence or instability property of the system through its linearized counterpart given by T. The separation of two infinitesimally close initial values $\mathbf{x}_0, \mathbf{x}_0'$ after time l is given approximately by

$$\mathbf{x}_l - \mathbf{x}_l' = M^l(\mathbf{x}_0) - M^l(\mathbf{x}_0') \approx \{T^l(\mathbf{x}_0)\}(\mathbf{x}_0 - \mathbf{x}_0'), \tag{1.3}$$

where M^l is the lth composition of M with itself, and $T^l(\mathbf{x}_0)$ denotes the Jacobian matrix $D_{M^l}(\mathbf{x}_0)$. By the chain rule of differentiation,

$$T^l(\mathbf{x}) = T(\mathbf{x}_{l-1}) \cdots T(\mathbf{x}_1)T(\mathbf{x}_0).$$

Denote the singular values of $T^l(\mathbf{x}_0)$ by $\delta_1(l, \mathbf{x}_0) \geq \delta_2(l, \mathbf{x}_0) \geq \cdots \delta_p(l, \mathbf{x}_0)$ (or alternatively, they are the square root of the eigenvalues of $\{T^l(\mathbf{x}_0)\}^T T^l(\mathbf{x}_0)$). The Lyapunov exponents, denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, are defined by

$$\lambda_i = \lim_{l\to\infty} \frac{1}{l} \log \delta_i(l, \mathbf{x}_0), i = 1, 2, \ldots, p. \tag{1.4}$$

The existence of above limits are guaranteed by the multiplicative ergodic theorem of Oseledec (1968) (see Section 2.4) for almost all $\mathbf{x}_0$ wrt an ergodic invariant measure ρ.

If $\lambda_1 > 0$, two trajectories from the linear system $\mathbf{y}_n = T(\mathbf{x}_{n-1})\mathbf{y}_n$ diverge exponentially fast for most choices of $\mathbf{y}_0$. It can be shown this implies that trajectories from the nonlinear system $\mathbf{x}_n = M(\mathbf{x}_{n-1})$ will diverge exponentially fast for any infinitesimally close initial conditions. Thus, $\lambda_1 > 0$ gives a quantification of the property of sensitive dependence on initial conditions. If for a typical initial value, a system has a bounded solution and at least one Lyapunov exponent, it is said to be chaotic. We will adopt this general definition of chaos. Though only discrete-time systems will be studied, analogous theory may be given for continuous-time systems, see Eckmann and Ruelle (1985), Ruelle (1989) for related discussions.

This definition of chaos in terms of instability can be generalized to a noisy system. For simplicity, we consider following stochastic difference equation

$$X_{n+1} = M(X_n) + GE_{n+1}, \tag{1.5}$$

where X_n, E_n are vectors in $\mathcal{R}^p$, and $M : \mathcal{R}^p \to \mathcal{R}^p$ is a differentiable map, and G is some constant $p \times p$ matrix. Furthermore, we assume that E_n's are iid with zero mean and covariance matrix Σ, and are independent of X_0. Under these assumptions, the sequence $\{X_n\}$ is a $p-$dimensional Markov chain. A crucial assumption we make is that the Markov chain (1.5) has an invariant measure ρ with compact support.

We can define product of Jacobian matrices:

$$T^l = T(X_{l-1}) \cdots T(X_1)T(X_0), \tag{1.6}$$

where $X_{l-1}, \ldots, X_1, X_0$ are part of trajectory from (1.5). Let $\delta_1(l) \geq \cdots \geq \delta_p(l)$ be the singular values of T^l. Oseledec's theorem (Section 2.4) insures that the following limits exist with probability one, for ρ-almost all X_0,

$$\lambda_i = \lim_{l \to \infty} \frac{1}{l} \log \delta_i(l), i = 1, 2, \ldots, p. \tag{1.7}$$

The quantities $\lambda_1 \geq \cdots \geq \lambda_p$ are called the Lyapunov exponents for noisy system (1.5). If ρ is ergodic, the λ_i's are constant. A noisy system which has bounded solution and at least one positive Lyapunov exponent is defined to be chaotic.

Determination of Lyapunov exponents can be based on simulating a typical trajectory using the algorithm e.g. in Parker and Chua (1989), if the system (1.5) is known. However, in practice, often the only available information is a scalar time series $\{x_t\}$, which represents observations from (1.5). Our focus is to determine the Lyapunov exponents for a time series $\{x_t\}$.

1.3 Chaotic Time Series

A scalar time series $\{x_t\}$ represents observations from a physical system, which usually has a multidimensional attractor. How can we hope to study the original system from analyzing the scalar signal $\{x_t\}$? It happens that this is possible if our focus is on determining the dynamical quantities of the original attractor. The state space reconstruction is the basis for recovering the properties of the original attractor from a scalar time series.

Specifically, given time series $\{x_t\}$, a standard way to reconstruct the state vectors is given by the delay coordinate vector, also called the *time delay* method,

$$X_t = (x_t, x_{t-\tau}, \ldots, x_{t-(p_e-1)\tau})^T, t = (p_e - 1)\tau + 1, \ldots,$$

where τ is the time delay, p_e is the embedding dimension. In practice, τ and p_e are often chosen by trial and error.

Usually, the desired quantity to be computed, such as a fractal dimension or the largest Lyapunov exponent, will stabilize and converge to a fixed value as p_e is increased beyond a certain threshold. A justification for this approach in the deterministic case is given by Takens' theorem (Takens 1981). Takens' theorem effectively says that the reconstructed state vectors will have the same dynamical properties of the original system if p_e is large enough. See Section 2.9 for more details.

Since a time series is usually subject to various sources of dynamical and measurement noises, we consider the following noisy model for our time series (without loss of generality we take $\tau = 1$):

$$x_{t+1} = m(x_t, \cdots, x_{t-p+1}) + \sigma\varepsilon_{t+1}, \tag{1.8}$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are iid random noises with zero mean and unit variance, and ε_{t+1} is independent of $x_t, x_{t-1}, \ldots$. This model is often called a nonlinear (pth-order) autoregressive model in time series, where m is called the autoregressive function. The time series comprises two parts: a low-dimensional deterministic system plus an infinite-dimensional random noise. Similar models are also considered in Smith (1992a), Nychka et al (1992).

It is easy to rewrite (1.8) in its state space form (1.5) by the time delay method. The Lyapunov exponents for the time series are defined based on its state space form. It is noted that the Jacobian matrix T for the state space form consists of the vector of partial derivatives of m as the first row, the next $(p-1)$ rows are given by the identity matrix and the zero vector of dimension $(p-1)$. However, in practice, p is often unknown. When the embedding dimension $p_e \geq p$, it can be shown that the first p Lyapunov exponents based on the reconstructed vectors are the same as those based on p and the remaining ones assume $-\infty$. In this sense, the Lyapunov exponents for a time series are defined consistently.

Since the Jacobian matrix involves only the partial derivatives of autoregression, only partial derivative estimation is needed in obtaining estimates of Jacobian matrices, from which estimates of Lyapunov exponents are derived. We propose that the locally weighted polynomial fit be used in the estimation of partial derivatives. This proposal is also consistent with Eckmann and Ruelle (1985) and Abarbanel et al (1991).

Nonlinear prediction in chaotic time series in deterministic context is studied in e.g. Farmer and Sidorowich (1987, 1988). The improvement of short-term forecasts may be seen as an effective way to distinguish chaos and nonlinearity in time series, see Sugihara and May (1990) and Casdagli (1992). The nonparametric regression estimation to be discussed below can be used for nonlinear prediction in a noisy time series, where the main issue is estimation of autoregression m based on past data.

1.4 Nonparametric Regression

Nonparametric regression has been a flexible tool of modeling dependence relationships between variables. We will be concerned only with the random design model, which is given by:

$$Y_i = m(X_i) + \nu^{1/2}(X_i)\varepsilon_i, \quad i = 1, \ldots, n \tag{1.9}$$

where X_i's are iid random variables in $\mathcal{R}^p$, ε_i's are iid scalar random variables with zero conditional mean and unit conditional variance, m is the mean regression function, and ν is the conditional variance.

The kernel method is the simplest and most widely used technique in nonparametric regression. A promising class of kernel estimators, usually called the locally weighted regression estimators has been studied in Fan(1993), Fan and Gijbels (1992), and Ruppert and Wand (1994). The locally weighted polynomial fit can be seen as a smooth version of the popular local polynomial fit, which is used e.g. in Stone (1977), Eckmann and Ruelle (1985), Abarbanel et al (1991), and Casdagli (1992). Other reasons for choosing the locally weighted polynomial fit include theoretical optimality considerations such as Fan (1993); no boundary effect, e.g. Fan and Gijbel (1992); less computation, particularly in multivariate data, e.g. Cleveland and Devlin (1988); and derivative estimation, e.g. Ruppert and Wand (1994).

The multivariate locally weighted polynomial fit in the nonparametric regression setup is studied in Chapter 3, where the two important cases, the *local linear fit* and the *local quadratic fit*, are investigated in detail. Explicit calculations on the asymptotic conditional bias and conditional covariance matrix of the regression and partial derivative estimators are given, generalizing the results of univariate derivative estimation in Ruppert and Wand (1994) to the multivariate case.

The results in Chapter 3 are then generalized to time series in Chapter 4. A related reference is Masry and Fan (1993), who have studied the univariate locally weighted polynomial fit for a mixing process. Asymptotic normality of the estimators will be established, along with the asymptotic bias and the asymptotic covariance matrix, under general short-range dependence conditions. We also discuss the extension of nonparametric estimation to fractal time series, that is, the finite-dimensional probability measure for the time series is only assumed to have a fractal dimension and may be singular with respect to the Lebesgue measure.

1.5 Estimation of Lyapunov Exponents

It is straightforward that estimates of the Lyapunov exponents can be defined based on Jacobian estimates. Let $\hat{T}^l$ be the estimated matrix product given as in (1.6) but with T's replaced by their estimates $\hat{T}$'s. Denote the singular values of $\hat{T}^l$ by $\hat{\delta}_1(l) \geq \cdots \geq \hat{\delta}_p(l)$. For any choice of l which depends on sample size n and tends to infinity as n does, the Lyapunov exponents estimators are given by

$$\hat{\lambda}_i(l) = \frac{1}{m} \log \hat{\delta}_i(l), i = 1, 2, \ldots, p, \text{ where } l = l(n).$$

There are at least two issues. One is numerical implementation. To insure numerical stability, the Lyapunov exponents should be computed using a proper algorithm, e.g. the QR algorithms as discussed in Eckmann and Ruelle (1985). See also Parker and Chua (1989). Another question is how to choose l and what is the associated convergence rate of $\lambda_i(l)$? It appears that l should be chosen as large as possible. However, it seems difficult to prove the convergence rates for such estimators. In the case of $\hat{\lambda}_1$, McCaffrey et al (1992) and Nychka et al (1992) have given a conjecture, concerning an estimator and the convergence rate. See also Ellner et al (1991). An additional problem may be the asymptotic distribution of the estimators.

To provide some insights for the above problems, we propose to study the estimators for fixed l. Obviously, the estimator $\hat{\lambda}_i(l)$ for fixed l converges (in some sense) to

$$\lambda_i(l) = \frac{1}{m} \log \delta_i(l),$$

for $i = 1, 2, \ldots, p$, where $\delta_1(l) \geq \cdots \delta_p(l)$ are the singular values of the true Jacobian product T^l.

We call $\lambda_1(l) \geq \cdots \geq \lambda_p(l)$ the $l-$step *local Lyapunov exponents* (for their dependence on the local trajectory $X_{l-1}, \ldots, X_1, X_0$). It should be pointed out that X_0 can be any point in the phase space. For any given fixed starting point $\mathbf{x}_0$, the interest may be in the behavior of $\lambda_i(l)$'s, which are functions of $E_1, \ldots, E_{l-1}$. For example, their distribution functions, means or the variances may be of interest. To answer these questions, it is important to study the estimators $\hat{\lambda}_i(l)$'s for any fixed $\mathbf{x}_{l-1}, \ldots, \mathbf{x}_1, \mathbf{x}_0$. We will show that the convergence rate of $\hat{\lambda}_i(l)$ for each i is the same as that of the partial derivative estimators. Furthermore, under general conditions, we will characterize the asymptotic distribution of the estimators.

It is noted that the problem of estimating the local Lyapunov exponents may be interesting itself for its more direct relevance to characterizing short-term predictability in time series. The local Lyapunov exponents provide measures of local divergence rates (corresponding to respective eigendirections). The largest local Lyapunov exponent is particularly important as it characterizes the largest possible divergence rate for each local trajectory. Quantifying the changes of the local divergence rate may have important implications for making forecasts, e.g. in weather forecasting, since we know the changes of speeds in phase space at which any errors will be amplified. The importance of the local Lyapunov exponents in time series analysis is also discussed in Wolff (1992) and Bailey (1993).

1.6 Organization of Chapters

The other chapters are organized as follows. Chaos and dynamical system theory, particularly the ergodic theory, is reviewed in Chapter 2. The main theory includes Lyapunov exponents, fractal dimensions, entropy, and Takens' embedding theorem. Chapter 3 is on multivariate locally weighted polynomial fit and partial derivative estimation. The nonparametric regression setup is considered, where the observations are assumed to be iid random vectors. In the time series context which is our main concern, the locally weighted polynomial fit is studied in Chapter 4. We also discuss the extension of nonparametric estimation to fractal time series, where the design probability measure is only assumed to have a fractal dimension. The application of partial derivative estimation to estimating Lyapunov exponents is considered in Chapter 5. We are able to establish the asymptotic distribution for the estimators of the local Lyapunov exponents. Some possible developments of the present work are discussed in Chapter 6.

Chapter 2

THEORY OF CHAOS

2.1 Introduction

A nonlinear dynamical system can generate orbits that are typically aperiodic, irregular, and apparently random, a phenomenon often called chaos. Chaos has provided an alternative way of interpreting irregular behavior in nature, particularly in physical sciences, e.g. turbulence in fluid mechanics. While a good understanding of the onset of chaos has been achieved using the geometric theory of dynamical systems, moderately excited chaotic systems require the ergodic theory of differentiable dynamical systems. The main subjects are the theory of dimensions, entropy, and Lyapunov exponents.

The theory of chaos has also been applied to economics, ecology, epidemiology, biology, etc., where typically only limited observations are available, and the data may be subject to various sources of noises, such as observational and dynamical noises (see Smith 1992a, Nychka et al 1992). Consequently, we will focus on noisy systems as our underlying models. The study of noisy chaos is still at its infant stage, so except in Section 2.5 where noisy systems are considered, we will mainly discuss the theory of deterministic chaos. Only discrete-time systems are discussed, since we will only be interested in modeling time series that has been obtained at equal time intervals. Analogous theory exists for continuous-time systems. See Eckmann and Ruelle (1985), Ruelle (1989).

The theory of dynamical systems and chaos are given in Devaney (1989), Farmer et al (1983), Guckenheimer and Holmes (1990), and Ott(1993). The ergodic theory of chaos is given in Eckmann and Ruelle (1985), Ruelle (1989), and Sinai (1989). We will review mainly the ergodic theory of chaos and will follow closely Eckmann and Ruelle (1985) and Ruelle (1989).

This chapter is organized as follows. Some well-known chaos examples are given in Section 2.2. Some basic concepts, such as dissipative systems, strange attractors, and asymptotic measures are defined in Section 2.3. In Section 2.4, we discuss the main theory of Lyapunov exponents. The important theorem, Oseledec's multiplicative ergodic theorem, is stated for a sequence of stationary random matrices. The Lyapunov exponents for a deterministic map or system are then defined, while the definition the Lyapunov exponents for a random map or stochastic dynamical system is postponed until Section 2.5, where a general stochastic dynamical system is defined and its ergodic properties are discussed.

Even though a simple system (refer to Section 2.2) still defies a mathematical analysis, some idealized systems, such as hyperbolic systems, Axiom-A systems, are well understood. Section 2.6 reviews some of this theory of dynamical systems, including the SRB measure.

The fractal dimensions remain the simplest concept, which are discussed in Section 2.7. Finite dimensionality of a dynamical system is an important property, and a fractal dimension gives a test of just that. The information-producing aspect of chaos, quantified by entropy, is discussed in Section 2.8. Section 2.9 reviews the state space reconstruction theory of time series, including Takens' theorem, and the invariance of dynamical properties.

2.2 Some Chaos Examples

This section gives some well-known examples that are known or believed to exhibit chaos. New concepts are often illustrated through these examples later on. These systems may also serve as benchmark examples as tests of statistical methods to be developed later on.

(a)**Logistic Map**

The logistic map is given by the quadratic function

$$m_\mu(x) = \mu x(1-x),$$

where the interest is on $\mu > 0, x \in I = [0,1]$. The difference equation given by m_μ has the form $x_{n+1} = m_\mu(x_n)$ or $x_n = m_\mu^n(x_0)$, where m_μ^n is the nth composition of m, and $x_0 \in I = [0,1]$.

The logistic map mimics the dynamics of population growth in biology and is well known to go through a whole spectrum of possible dynamics as μ varies over $[0,4]$ (Devaney 1989). This is best illustrated by the *bifurcation diagram*, see Figure 2.1 a,b, where the set $\{m^i(x_0)\}_{i=N_0}^N$ for a typical initial value x_0 is plotted against the control parameter μ. Here, $x_0 = .2, N_0 = 200, N = 400$. Figure 2.1a corresponds parameter μ between 2.5 and 4.0. Figure 2.1b is a refinement of one section in Figure 2.1a, corresponding to μ between 3.84 and 3.86.

The diagram visualizes the change of the dynamical behavior of m_μ as the control parameter μ varies (occurrence of bifurcation). It is seen that,when $\mu < 3$ the logistic map has simple dynamics, since it can be shown easily that every initial value in (0,1) is attracted to the fixed point $p_\mu = (\mu - 1)/\mu$. As μ passes through 3, the dynamics of the map becomes slightly more complicated: a new periodic point of period 2 is born. The scenario of period doubling occurs as μ tends to $\mu_\infty \approx 3.570$, at which the orbit converges to a set, the Feigenbaum attractor, which has the structure of the Cantor set, in the sense that it is a closed, totally disconnected and perfect subset (Devaney 1989).

At $\mu \in (\mu_\infty, 4)$ several types of behaviour occur. There are several windows where the map has periodic orbit. However it is believed that the set $\{\mu : \lambda_\mu > 0\}$, where λ_μ is the Lyapunov exponent (to be defined later) for the map m_μ, i.e. the set of μ's at which the map is chaotic(and furthermore its attractor has an absolutely continuous invariant probability measure) has positive Lebesgue. However the dynamics in this region are not completely understood yet! See e.g. Sinai (1989). It can be shown rigorously that $m_4(x) = 4x(1-x)$ is chaotic on the interval $I = [0, 1]$. Furthermore, the invariant measure has density $f(x) = 1/\pi\sqrt{x(1-x)}$.

The logistic map is noninvertible. This map is typical of the one-dimensional chaos, which occurs only for a noninvertible map. In the case of a diffeomorphism M, i.e. the map is onto and one-to-one, both M and M^{-1} are differentiable, chaos occurs only in two or more dimensions (Ruelle 1989). In higher-dimensional map, a new object is created, so-called the attractor, which describes geometrically the long-term behavior of a system. The Hénon mapping is a famous example of a two-dimensional mapping that is believed to exhibit chaos.

(b)**Hénon mapping**

The difference equations from the map are given by:

$$\begin{aligned} x_{t+1} &= by_t + 1 - ax_t^2 \\ y_{t+1} &= x_t, \end{aligned}$$

where a and b are constants. Values $a = 1.4$ and $b = 0.3$ are often chosen for study (Hénon, 1976).

The evidence of chaos for the Hénon mapping with above chosen parameters can be seen through numerical study. The Hénon attractor is the scatter plot of the orbits $\{(x_t, y_t)\}_{n=N_0}^{N}$ for some typical initial value (x_0, y_0), where N_0 is the number of the transient steps that are discarded. See Figure 2.2, where $a = 1.4, b = 0.3, x_0 = .5, y_0 = .6, N_0 = 500, N = 30000 + N_0$. The Hénon attractor is locally a product of a line segment and a Cantor set. Another example is the Ikeda map, which is defined in the complex z plane.

(c)**Ikeda map**

$$z(n+1) = p + Bz(n)\exp\{ik - i\alpha/[1+|z(n)|^2]\},$$

where parameter values $p = 1.0, B = 0.9, k = 0.4$, and $\alpha = 6.0$ are often chosen. The Ikeda attractor is given in Figure 2.3, a plot of $\{Z(n)\}_{n=N_0}^{N}$, where the above parameter values are used, $N_0 = 500, N = 10000 + N_0, z(0) = 21.2 + 12.5i$.

The above examples are discrete-time dynamical systems given by difference equations or mappings. In contrast, continuous-time dynamical systems are given by differential equations. A celebrated example is the Lorenz system (Lorenz 1963).

(d)**Lorenz System**

It is given by the autonomous ordinary differential equations in $\mathcal{R}^3$.

$$\begin{aligned}\frac{dx(t)}{dt} &= -\sigma x(t) + \sigma y(t)\\ \frac{dy(t)}{dt} &= -x(t)z(t) + rx(t) - y(t)\\ \frac{dz(t)}{dt} &= x(t)y(t) - bz(t)\end{aligned}$$

where values $\sigma = 10, b = 8/3, r = 28$ are often chosen for numerical study (Lorenz (1963)).

The Lorenz system is one of the mostly studied systems. See Guckenheimer and Holmes (1990), Sparrow (1982). The Lorenz equations are obtained by truncation of the Navier-Stokes equations, and give an approximate description of a horizontal fluid layer heated from below. This is similar to what happens in the earth's atmosphere. For sufficiently intense heating (represented by r) the time evolution has sensitive dependence on initial conditions, thus representing a very irregular and chaotic convection. This fact was used by Lorenz to justify the so-called "butterfly effect", a metaphor of the imprecision of weather forecasting. The Lorenz system is the first example that is derived from an actual physical process and gives rise to a type of attractor which is nonclassical (neither periodic nor quasiperiodic). See Figure 2.4a,b. The Euler algorithm is used for numerical integration, with initial values

$x(0) = 4, y(0) = 10, z(0) = 8$, step size $\Delta = .01$, warm up steps $N_0 = 2000$, plotted steps $N = 20000$ which corresponds to time interval of length $T = 200$. Details of the algorithm is given in Parker and Chua (1989), where more sophisticated and accurate algorithms are also available.

Another continuous-time system, which is often studied in the literature for its simplicity, is the Rössler system.

(e)**Rössler System**

It is given by the autonomous ordinary differential equations in $\mathcal{R}^3$.

$$\begin{aligned} \frac{dx(t)}{dt} &= -(y(t) + z(t)) \\ \frac{dy(t)}{dt} &= x(t) + ay(t) \\ \frac{dz(t)}{dt} &= b + x(t)z(t) - cz(t) \end{aligned}$$

where a, b, c are parameters to be chosen. Parameter values $a = .15, b = .2, c = 10$ are often chosen, see McCaffrey et al (1992). The Rössler attractor is given in Figure 2.5a,b. The Euler algorithm is used, where $x(0) = 5, y(0) = 8, z(0) = 9$, step size $\Delta = 0.01$, warm up steps $N_0 = 5000$, plotted steps $N = 30000$ corresponding to time interval of length $T = 300$.

2.3 Strange Attractor

Strange attractor. A discrete dynamical system is given by iterated maps $\mathbf{x}_n = M^n \mathbf{x}_0$, where $\mathbf{x}_0$ is an initial value in a set D, $M : D \rightarrow D$ is a map satisfying $M^0 \mathbf{x} \equiv \mathbf{x}$ and $M^{n_1+n_2} = M^{n_1} \circ M^{n_2}$, where $\circ$ denotes composition. D is usually a subset of R^p but can be quite general. D is called the state space or phase space. In the following we will be mainly concerned with dissipative systems, for which the volume in the phase space is usually contracted. The Hénon map, Ikeda map, Lorenz and Rössler systems are all dissipative systems.

For a dissipative system, one can generally assume that there is a set U in $\mathcal{R}^p$ which is contracted by time evolution asymptotically to a compact set $A = \cap_{n \geq 0} M^n U$. The limit set A is clearly an invariant set, in the sense that $M^n A = A$, for all n, and it has zero volume as a result of contraction.

The set A is said to be *attracting* with fundamental neighborhood U, if for every open set $V \supset A$ we have $M^n U \subset V$ for all sufficiently large n. Its basin of attraction is defined to be the set of initial points $\mathbf{x}$ such that $M^n \mathbf{x}$ approaches A as $n \to \infty$, i.e. $\cup_{n<0} M^n U$.

An important concept is the attractor, named after the fact that it is an attracting set. We will not attempt to give a mathematical definition of an attractor. See Ruelle (1989) for more details. Operationally, it is a set on which experimental points M^n accumulate for large n for most initial values (i.e. attracting). An attractor A should be invariant, but need not have an open basin of attraction. An attractor should also satisfy irreducibility, i.e. there exists a point $\mathbf{x}_0 \in A$ such that for each $\mathbf{x} \in A$ there is a positive n such that $M^n \mathbf{x}_0$ is arbitrary close to $\mathbf{x}$ (topological transitivity). Equivalently, there exists a dense orbit $\{M^n \mathbf{x}_0\}_{n \geq 0}$ in A.

An attractor is called a *strange attractor* if it has the property of “sensitive dependence on initial conditions”, i.e. the evolutions from two infinitesimally close initial values will diverge exponentially fast. This property is related to the dynamical properties of an attractor, not just to its geometry. A dynamical system is called chaotic if it has a strange attractor.

For example, the Hénon and Ikeda attractors from discrete-time systems are believed to be strange, and Lorenz and Rössler attractors from continuous-time systems appear to be strange. None of the claims has yet been proved, but numerical study shows that they are strange or chaotic, that is, they all have one positive Lyapunov exponent.

The presence of chaos implies a strong sensitivity to small fluctuations. In a computer study, due to roundoff error, the simulated trajectory may be completely different from the true one. How to interpret it is clearly an important question. This is also important if we are interested in the study of situations where the systems are always affected by some noise. We expect that the properties which are investigated should be stable under small fluctuations. So we require that an attractor should have stability under small random perturbations, or more precisely the asymptotic measure ρ should be stable under such small perturbations (see below).

Invariant measure. An attractor A gives a global picture (a geometrical property) of the long-term behavior of a dynamical system. More details about an attractor are given by a probability measure ρ on A, which describes how frequently various parts of the attractor are visited by the orbit. We will review some concepts in ergodic theory, and will use the abstract setup. Given a probability space $(\mathcal{X}, \mathcal{A})$, M is a measurable transformation, ρ is a probability measure. ρ is called invariant wrt M, if $\rho(A) = \rho(M^{-1}A)$, for any $A \in \mathcal{A}$.

The set A is an invariant set if $M^{-1}A = A$. The measure ρ is *ergodic* if $\rho(A) = 0$ or 1 for every invariant set $A \in \mathcal{A}$. The existence of an invariant probability measure on the attractor is guaranteed in the following situation: if A is a compact invariant set wrt M, there exists an invariant probability measure ρ with support contained in A. Moreover, ρ can be chosen to be ergodic.

The property of ergodicity enables one to study an invariant ergodic measure on an attractor from one typical orbit, as shown by the ergodic theorem.

Theorem 2.1 (Ergodic Theorem) *Given a probability space $(\mathcal{X}, \mathcal{A}, \rho)$, and a measure-preserving map M, i.e. $\rho(A) = \rho(M^{-1}A)$, for any integrable function Φ, the following limit*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \Phi(M^i \mathbf{x}_0) = \bar{\Phi}(\mathbf{x}_0) \tag{2.1}$$

exists for ρ almost all $\mathbf{x}_0$. If ρ is further ergodic, then $\bar{\Phi}$ is constant and $\bar{\Phi} = E\Phi(\mathbf{x}_0)$.

An important case in Theorem 2.1 is to choose $\Phi = 1_A$ for $A \in \mathcal{A}$. If ρ is ergodic, equation (2.1) becomes:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1_{\{M^i \mathbf{x}_0 \in A\}} = \rho(A) \ \text{ a.s..} \tag{2.2}$$

Notice that the left-hand sides of equations (2.1) and equation (2.2) are time averages, while the right-hand sides are one-time phase averages, i.e. averages over the phase space weighted according to the probability ρ. The ergodic theorem says "time average=phase average" for an ergodic invariant probability measure. In particular in the case of equation (2.2), the right-hand side is the average time spent by the dynamical system in A and equation (2.2) says that the invariant probability measure can be obtained from one complete orbit from a typical initial value under the ergodicity assumption. This implies that it is in principle possible for us to do statistical analysis of an asymptotic measure by analyzing a typical orbit.

However, ρ may be singular with respect to the Lebesgue measure, so that the initial values for which the orbits can generate ρ may be in a set of Lebesgue measure zero. Thus, a singular ρ may not be realizable in practice. Furthermore, in typical cases there are uncountably many distinct ergodic measures.

Example Consider the map $m : [0,1) \rightarrow [0,1)$ defined by $m(x) = 2x(\text{mod } 1)$. Each number in [0,1) has a binary expansion $0.a_1a_2a_3\cdots$, where for each $i, a_i = 0$ or 1. Clearly, f is a shift replacing $0.a_1a_2a_3\cdots$ by $0.a_2a_3\cdots$. For any given value r between 0 and 1, a probability distribution ρ_r can be defined by requiring that a_i be 0 with probability r, and 1 with probability $1-r$ (independently for each i). Then ρ_r is invariant under the shift, and in fact ergodic. There are uncountably many such measures, corresponding to the different values of r in (0,1).

The invariant measure on an attractor A given by the time averages of an orbit is called an asymptotic measure, which describes how frequently various parts of A are visited by the orbit. A natural choice of an asymptotic measure for a dynamical system is the one given by the time averages for typical initial values, such as those in a set of positive Lebesgue measure. For example, if ρ is absolutely continuous wrt the Lebesgue measure, it is our choice because it is given by initial values in a positive Lebesgue set and will be easily observed. The *SRB measure* (given in Section 2.6) satisfies this selection condition.

Going back to above example, the measure $\rho_{0.5}$ corresponding to the Lebesgue measure l is the natural choice because it is given by the time averages for almost all initial values wrt the Lebesgue measure. It can be shown that all other measures $\rho_r, r \neq 0.5$ are singular with respect to l. Another example is given by the logistic map $m(x) = 4x(1-x)$. A natural measure is given by the beta density $f(x) = 1/\{\pi\sqrt{x(1-x)}\}$.

Another selection process is related to the requirement of stability of the measure under random perturbation. A physical system is often subject to some random noises at a small level σ, and it can be considered as a Markov process. In a computer study, roundoff errors play the role of the random noises. A Markov process often has only one stationary measure ρ_σ, and we may hope that ρ_σ tends to a specific measure, often called the Kolmogorov measure when the noise level $\sigma \to 0$ (Eckmann and Ruelle 1985).

These proposals can be substantiated in some simple systems, such as Axiom A systems which are defined in Section 2.6.

2.4 Lyapunov Exponents

In this section, we will justify the existence of Lyapunov exponents introduced in Chapter 1. We will first define Lyapunov exponents in the general context of a stationary sequence of random matrices, then define them for a deterministic system. The Lyapunov exponents can also be defined rigorously for a stochastic dynamical system, as will be done in Section 2.5.

Consider a sequence of stationary $p \times p$ matrices $A_1, A_2, \ldots$. Denote $\log^+ x = \max\{0, \log x\}$, where $x \geq 0$, we assume that

$$E \log^+ \|A_1\| < \infty.$$

The sequence is said to be ergodic if all sets in its tail $\sigma-$field $\mathcal{F}_\infty$ have probability 0 or 1, where $\mathcal{F}_\infty = \cap_n \mathcal{F}_n, \mathcal{F}_n = \sigma\{A_n, A_{n+1}, \ldots\}$.

Define the product of matrices:

$$T^n = A_n \cdots A_2 A_1.$$

We will be concerned with characterizing the growth rates of T^n as n grows. The following theorem on the products of random matrices was first proved by Oseledec (1968), who generalized an earlier result on the norm of products of stationary random matrices (maximum growth rate) due to Furstenberg and Kesten (1960). We state the theorem in the general context of a stationary sequence of random matrices $\{A_i\}$.

Theorem 2.2 (Multiplicative ergodic theorem) *With probability 1, the following limit exists:*

$$\lim_{n\to\infty} \{(T^n)^T T^n\}^{1/(2n)} = \Lambda \tag{2.3}$$

(where $(T^n)^T$ is the transpose of T^n.)

The logarithms of the eigenvalues of Λ are called Lyapunov exponents, *denoted by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. If the sequence $\{A_i\}$ is also ergodic, the Lyapunov exponents are constant, though Λ may be random in general. Let $\lambda^{(1)} > \lambda^{(2)} > \ldots$ be the distinct Lyapunov exponents, and $d^{(i)}$ the multiplicity of $\lambda^{(i)}$. Let $E^{(i)}$ be the subspace of $\mathcal{R}^p$ spanned by the eigenvectors associated with eigenvalues $\leq \exp\lambda^{(i)}$ of Λ (the $E^{(i)}$'s are random in general). Then $R^p = E^{(1)} \supset E^{(2)} \supset \cdots$ and the following holds, with probability 1,*

$$\lim_{n\to\infty} \frac{1}{n} \log \|T^n u\| = \lambda^{(i)}, \quad \textit{if } u \in E^{(i)} \backslash E^{(i+1)}. \tag{2.4}$$

In particular, for all vectors u that are not in the subspace $E^{(2)}$ (thus for Lebesgue measure almost all u), the limit of left-hand side of Equation (2.4) is the largest Lyapunov exponent.

Lyapunov exponents for deterministic systems. Consider a discrete-time dynamical system on $\mathcal{R}^p$ given by:

$$\mathbf{x}_{n+1} = M(\mathbf{x}_n) \tag{2.5}$$

where $M : \mathcal{R}^p \to \mathcal{R}^p$ is a differentiable map. We denote by $T(\mathbf{x})$ the $p \times p$ Jacobian matrix of M at $\mathbf{x}$. The Jacobian matrix for the nth iterate M^n is given by the chain rule:

$$T_{\mathbf{x}}^n = T(M^{n-1}\mathbf{x}) \cdots T(M\mathbf{x})T(\mathbf{x}).$$

Now, assuming that there exists an invariant measure ρ wrt M with compact support, the sequence of random matrices $\{T(M^{n-1}\mathbf{x})\}$ is stationary and satisfies the conditions of Theorem 2.2, given $\mathbf{x}$ being randomly distributed according to ρ. So we have the following theorem for a deterministic system M.

Theorem 2.3 (Lyapunov exponents in dynamical system) *The limit*

$$\lim_{n\to\infty} \{(T_{\mathbf{x}}^n)^T T_{\mathbf{x}}^n\}^{1/(2n)} = \Lambda_{\mathbf{x}} \tag{2.6}$$

exists for ρ-almost all $\mathbf{x}$. The logarithms of the eigenvalues of $\Lambda_{\mathbf{x}}$ are called Lyapunov exponents, *denoted by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. If ρ is further ergodic, the Lyapunov exponents are independent of $\mathbf{x}$.*

Furthermore, let $\lambda^{(1)} > \lambda^{(2)} > \ldots$ be the distinct Lyapunov exponents, and $d^{(i)}$ the multiplicity of $\lambda^{(i)}$. Let $E_{\mathbf{x}}^{(i)}$ be the subspace of $\mathcal{R}^p$ spanned by the eigenvectors associated with eigenvalues $\leq \exp \lambda^{(i)}$ of Λ. Then $R^p = E_{\mathbf{x}}^{(1)} \supset E_{\mathbf{x}}^{(2)} \supset \cdots$ and the following holds, with probability 1,

$$\lim_{n\to\infty} \frac{1}{n} \log \|T_{\mathbf{x}}^n u\| = \lambda^{(i)}, \quad \textit{if } u \in E_{\mathbf{x}}^{(i)} \backslash E_{\mathbf{x}}^{(i+1)}. \tag{2.7}$$

In particular, for all vectors u that are not in the subspace $E_{\mathbf{x}}^{(2)}$ (thus for Lebesgue measure almost all u), the limit of left-hand side of Equation (2.7) is the largest Lyapunov exponent.

Intuitively, if there is a small change $\delta\mathbf{x}_0$ in the initial value $\mathbf{x}$, the corresponding change at time n is given by:

$$\begin{aligned} \delta\mathbf{x}_n &= T_{\mathbf{x}}^n \delta\mathbf{x}_0 \\ &= T(M^{n-1}\mathbf{x}) \cdots T(M\mathbf{x})T(\mathbf{x})\delta\mathbf{x}_0. \end{aligned}$$

For most $\delta\mathbf{x}_0$, i.e. for $\delta\mathbf{x}_0 \notin E_{\mathbf{x}}^{(2)}$, we have $\delta\mathbf{x}(n) \approx \delta\mathbf{x}_0 e^{n\lambda_1}$. So $\lambda_1 > 0$ corresponds to exponential divergence. In the case of a bounded attractor which is our main interest, if $\delta\mathbf{x}_0$ is finite rather than infinitesimally small, the growth of $\delta\mathbf{x}_n$ will not go on infinitely, since $\delta\mathbf{x}_n$ cannot be larger than the diameter of the attractor. Actually, what we often observe is that, the initial exponential growth is rapidly followed by folding and contracting.

Next, we discuss a stochastic dynamical system, and show how Lyapunov exponents can be defined for such a system.

2.5 Noisy Systems

A finite-dimensional deterministic system may not be attainable in practical modeling due to small random noises, such as computer roundoff error, external noise from the environment, measurement error, or even lack of fit due to some high-dimensional component in the system which is impossible to estimate with limited data. More discussions are given in Ruelle (1989), Smith (1992a,b), Casdagli (1992), Nychka et al (1992), Chan and Tong (1994), and Lasota and Mackey (1994). Consequently, we will consider a stochastic dynamical system as our underlying model. This viewpoint appears to bring us closer to traditional stochastic modeling framework, such as that discussed e.g. in Jazwinski (1970), Tong (1990). However, the differences may be more important. The present approach emphasizes the finite-dimensional deterministic structure, and deals with the case when the random noises (high to infinite-dimensional component) are small enough and play a minor role.

A general vector stochastic difference equation (in discrete-time) is given by:

$$X_{n+1} = M(X_n) + G(X_n)E_{n+1}, \tag{2.8}$$

where $M(\cdot)$ is usually a differentiable map in $\mathcal{R}^p$, G is a $p\times p$ matrix whose components are $p-$variate functions, E_n's are assumed to be iid $p-$dimensional dynamical noises with zero mean, and covariance matrix Σ. Further, the E_n's are independent of $\mathbf{x}_0$.

As a consequence, E_n is independent of $X_{n-1}, \ldots, X_1, X_0$ for any n, and the sequence $\{X_n\}$ is a $p-$dimensional Markov chain.

We will assume that the Markov chain $\{\mathbf{x}_n\}$ has an invariant and ergodic measure ρ which has compact support. Jusfication of this assumption may be provided by the Markov chain theory as given e.g. Nummelin (1984). A progress in this direction has been made in Chan and Tong (1994).

We have given a discrete-time stochastic dynamical system. Continuous-time dynamical systems can also be given which involves more advanced stochastic differential equation theory, see e.g. Lasota and Mackey (1994). Equation (2.8) may be seen as the system equation. An observation model may also be given which may include the measurement noises. We will not go further into this framework, see e.g. Smith (1992a), Jazwinski (1970) for more details.

Now we justify the definition given in Chapter 1 of Lyapunov exponents for a noisy system. Consider model (2.8), where for simplicity we take G to be a constant matrix. Denote $T(\mathbf{x})$ as the tangent map of M evaluated at $\mathbf{x}$, and denote the product of n random matrices by

$$T^n = T(X_{n-1}) \cdots T(X_1)T(X_0).$$

One motivation for considering T^n is that, the system (2.8) starting at X_0 and $X_0' = X_0 + \delta X_0 (\delta X_0$ small), for which we assume same noise sequence $E_1, E_2, \ldots$, the separation at the nth step is given by $\delta X_n = X_n' - X_n = T^n \delta X_0$.

If X_0 is chosen according to ρ, the sequence $\{X_n\}_{\{n \geq 0\}}$ is stationary and ergodic. So the sequence of random matrices $\{T(X_n)_{\{n \geq 0\}}\}$ is stationary and ergodic, and the conditions of Theorem 2.2 are satisfied. The Lyapunov exponents for $\{T(X_n)\}_{\{n \geq 0\}}$ can thus be defined and are called the *Lyapunov exponents for a noisy system* (2.8).

From another viewpoint, equation (2.8) may be regarded as generated from compositions of independent random maps

$$X_n = H_n \circ \cdots \circ H_2 \circ H_1,$$

where $H_i(\cdot) = M(\cdot) + G(\cdot)E_i, i = 1, 2, \ldots,$ and M, G are fixed as defined before, but E_i's are iid. So we may say that the Lyapunov exponents are defined for the random map H. Lyapunov exponents for a general random map can also be defined, see e.g. Kifer (1986). A stochastic dynamical system or a random map with at least one positive Lyapunov exponent is said to be *chaotic*. In this work, what we mean by chaos or noisy chaos should be interpreted in a noisy system in above sense.

2.6 Some Advanced Theory

Now we go back to deterministic systems. We will state some mathematical results for some idealized systems. It is noted that properties of some simple systems, e.g. in Section 2.2, remain to be proved.

We need some basic concepts. We say that M is a C^0-diffeomorphism or homeomorphism if M is one-to-one, onto, and continuous, and M^{-1} is also continuous. We say that $M : D \to D$ is a C^r- diffeomorphism, if M is one-to-one, onto, and both M and its inverse M^{-1} are $r-$times continuously differentiable. We will use the word "smooth" and diffeomorphism to mean C^1, unless stated otherwise. A smooth $d-$dimensional manifold $D \subset \mathcal{R}^p$ is a set for which each $\mathbf{x} \in D$ has a neighborhood U for which there is a smooth invertible mapping (diffeomorphism) $\Phi : \mathcal{R}^d \to U (d \leq p)$. Intuitively speaking, a $d-$dimensional manifold looks locally like $\mathcal{R}^d$. We say that a subset D in $\mathcal{R}^p$ is compact if its every open cover has a finite subcover.

In dynamical system theory, an important concept is hyperbolicity, which plays a crucial role in the development of much of the mathematical theory. It is easy to explain it in case of a fixed or a periodic point $\mathbf{x}$. A fixed point (or periodic point of period $n,$)$\mathbf{x}$ is said to be *hyperbolic* when $D_M(\mathbf{x})$ (or $D_{M^n}(\mathbf{x})$) has no eigenvalues with zero real part.

For a fixed point $\mathbf{x}$, the stable subspace $E^s_{\mathbf{x}}$ of the linear map $D_M(\mathbf{x})$ is the span of n_s eigenvectors whose eigenvalues have modulus< 1. The unstable subspace $E^u_{\mathbf{x}}$ is the span of n_u eigenvectors whose eigenvalues have modulus > 1. The center subspace $E^c_{\mathbf{x}}$ is the span of n_c eigenvectors whose eigenvalues have modulus $= 1$. Consider the system from linear map $T = D_M(\mathbf{x})$, i.e. $\mathbf{y}_n = T^n \mathbf{y}_0$. The orbits $\{\mathbf{y}_n\}$ in $E^s_{\mathbf{x}}$ and $E^u_{\mathbf{x}}$ are characterized by contraction and expansion, respectively. Hyperbolicity implies that the center space does not exist. Similar statements can be said about the linear map $D_{M^n}(\mathbf{x})$ in case of a periodic point $\mathbf{x}$.

Nonlinear analogues of linear spaces $E^s_{\mathbf{x}}$ and $E^u_{\mathbf{x}}$ can be defined for nonlinear M, which are called the invariant *stable* and *unstable manifold*, respectively. In the case of a fixed point $\mathbf{x}$, the stable manifold, denoted by $W^s_{\mathbf{x}}$, is a smooth invariant curve (in general a manifold), with dimension equal to that of $E^s_{\mathbf{x}}$, passing through $\mathbf{x}$ tangent to $E^s_{\mathbf{x}}$, and composed of points $\mathbf{y}$ such that $d(M^n\mathbf{y}, \mathbf{x}) \to 0$ when $n \to +\infty$. The *unstable manifold* can be defined in the same way, replacing n with $-n$ in the definition. Similar definition can be given for a periodic point.

In general, given an invariant ergodic measure ρ for M, for ρ-almost all point $\mathbf{x}$, the multiplicative ergodic theorem asserts the existence of linear spaces $E^{(1)}_{\mathbf{x}} \supset E^{(2)}_{\mathbf{x}} \supset \cdots$ such that

$$\lim_{n\to\infty} \frac{1}{n} \log \|T^n_{\mathbf{x}} u\| \le \lambda^{(i)}, \text{ if } u \in E^{(i)}_{\mathbf{x}}.$$

The exponential expansion or contraction for the linear system T has important implication for the nonlinear M. Actually, one can define a nonlinear analog of $E^{(i)}_{\mathbf{x}}$. If $\lambda^{(i)} < 0$, one can define global stable manifolds as

$$W^{(i)s}_{\mathbf{x}} = \{\mathbf{y} : \lim_{n\to\infty} \frac{1}{n} \log d(M^n\mathbf{x}, M^n\mathbf{y}) \le \lambda^{(i)}\}.$$

These global manifolds, though locally smooth, tend to fold and accumulate in a very complicated manner. We can also define the stable manifold of $\mathbf{x}$ by

$$W^s_{\mathbf{x}} = \{\mathbf{y} : \lim_{n\to\infty} \frac{1}{n} \log d(M^n\mathbf{x}, M^n\mathbf{y}) < 0\}.$$

(It is the largest of the stable manifolds, equal to $W^{(i)s}_{\mathbf{x}}$ where $\lambda^{(i)}$ is the largest negative Lyapunov exponent.) The unstable manifolds $W^u_{\mathbf{x}}$ can be defined simply through replacing n by $-n$ in above definitions. Eckmann and Ruelle (1985) give more details.

A point $\mathbf{x} \in \mathcal{R}^p$ is *wandering* if it has a neighborhood U such that $M^n(U) \cap U = \emptyset$ for n large enough. The *nonwandering* set $\Omega(M)$ is the set of all points which are not wandering. The $\Omega(f)$ is a compact $M-$invariant subset of $\mathcal{R}^p$, and contains the attractor of the dynamical system.

An $M-$invariant set A is *hyperbolic* if there exists a continuous invariant direct sum decomposition $T_A\mathcal{R}^p = E_A^u \oplus E_A^s$ with the property that there are constants $C > 0$ and $0 < \lambda < 1$ such that:
(1) if $\nu \in E_{\mathbf{x}}^u$, then $|D_M^{-n}(\mathbf{x})\nu| \leq C\lambda^n|\nu|$;
(2) if $\nu \in E_{\mathbf{x}}^s$, then $|D_M^n(\mathbf{x})\nu| \leq C\lambda^n|\nu|$.
See Eckmann and Ruelle (1985), Guckenheimer and Holmes (1990).

If the whole compact manifold D is hyperbolic, M is called an Anosov diffeomorphism, e.g. Thom's toral automorphisms and Arnold's cat map in Eckmann and Ruelle (1985), and Devaney (1989) are Anosov diffeomorphisms. If $\Omega(M)$ is hyperbolic, and the periodic points of M are dense in $\Omega(M)$, M is called an Axiom-A diffeomorphism. Every Anosov diffeomorphism is an Axiom-A diffeomorphism. Other examples include Smale's horseshoe, and the solenoid. See Devaney (1989), Guckenheimer and Holmes (1990) for details.

We now introduce the concept of *structural stability.* We will follow Ruelle (1989). Two diffeomorphisms $S, M : D \rightarrow D$ are topologically conjugate if there exists a homeomorphism Φ such that $\Phi \circ S = M \circ \Phi$. Then, a diffeomorphism M of a compact manifold D is structurally stable if it has a neighborhood V in the C^1 topology (i.e. M and S are close if both the diffeomorphisms and their (partial) derivatives are close), such that every S in V is topologically conjugate to M.

The concept of structural stability is completely different from that of dynamical stability (Lyapunov stability). The latter refers to individual orbits and requires that there is no sensitive dependence on initial conditions. The former refers to the whole system and asks that, under small C^1 perturbations of the system, the qualitative features of the system are preserved.

It is known that if M satisfies Axiom-A, together with a hypothesis of *strong transversality*, it is structurally stable. Strong transversality means that every intersection point of $W^s_{\mathbf{x}}$ and $W^u_{\mathbf{x}}$ is transversal for all $\mathbf{x}, \mathbf{y} \in \Omega(M)$ ($W^s_{\mathbf{x}}$ and $W^u_{\mathbf{x}}$ intersect transversally in a point $\mathbf{z} = W^s_{\mathbf{x}} \cap W^u_{\mathbf{x}}$, if the tangent spaces $T_{\mathbf{x}} W^s_{\mathbf{x}}, T_{\mathbf{x}} W^u_{\mathbf{x}}$ span $T_{\mathbf{x}} \mathcal{R}^p$).

Now the SRB measure can be defined. Intuitively, due to the stretching in the unstable direction, we may expect that an invariant measure is smooth along the unstable directions. Such a measure is called an SRB measure. More precisely, consider a ρ-measurable set of the form $S = \bigcup_{\alpha \in A} S_\alpha$, where the S_α are disjoint pieces of unstable manifolds W^u(each S_α can be constructed by the intersection of a local unstable manifold with S). If this decomposition is ρ measurable, then one has

$$\rho \text{ restricted to } S = \int \rho_\alpha l(d\alpha),$$

where l is a measure on A, and ρ_α is a probability measure on S_α, a conditional probability measure associated with the decomposition $S = \bigcup_{\alpha \in A} S_\alpha$. The ρ_α are defined $l-$almost everywhere. The ergodic measure ρ is an *SRB measure* (for Sinai-Ruelle-Bowen) if its conditional probabilities ρ_α are absolutely continuous with respect to Lebesgue measure for some choice of S and its decomposition.

The SRB measure turns out to be the natural choice in many cases. It can be shown that the time average of an orbit (Equation (2.2)) in section 2.3 tends to the SRB measure ρ when $n \to \infty$, not just for $\rho-$almost all $\mathbf{x}$, but for $\mathbf{x}$ in a set of positive Lebesgue measure. We have the following result on SRB measures for Axiom-A systems (refer to Eckmann and Ruelle (1985)). Consider a dynamical system determined by a twice differentiable diffeomorphism M on a p-dimensional manifold D. Suppose that A is an Axiom-A attractor, with basin of attraction U.
(a) There is one and only one SRB measure with support in A.
(b) There is a set $S \subset U$ such that $U \setminus S$ has zero Lebesgue measure, and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \delta_{M^i \mathbf{x}} = \rho \text{ whenever } \mathbf{x} \in S.$$

The following result due to Pugh and Shub (1984) given in Eckmann and Ruelle

(1985) shows that the requirement of Axiom A can be replaced by weaker information about the Lyapunov exponents. Let M be a twice differentiable diffeomorphism of an $p-$dimensional manifold D and ρ an SRB measure such that all Lyapunov exponents are different from zero. Then there is a set $S \subset D$ with positive Lebesgue measure such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \delta_{M^i \mathbf{x}} = \rho$$

for all $\mathbf{x} \in S$. (P.640 of Eckmann and Ruelle (1985).)

It is worth mentioning a shadowing property for a hyperbolic invariant set. One says that a sequence $\mathbf{x} = \{x_i\}_{i=a}^{b}$ is an *α-pseudo-orbit* for M if $d(x_{i+1}, M(x_i)) < \alpha$ for all $a \leq i < b$. A point y *β-shadows* $\mathbf{x}$ if $d(M^i(y), x_i) < \beta$ for $a \leq i < b$. If A is a hyperbolic invariant set, then for every $\beta > 0$, there is an $\alpha > 0$ such that every α-pseudo-orbit $\{x_i\}_{i=a}^{b}$ in A is β-shadowed by a point $y \in A$. This result is known as Bowen's shadowing lemma, see p.251 of Guckenheimer and Holmes(1990). It implies that, while a computer may not calculate the exact orbit, what it does find is nevertheless an approximation to some true orbit of the system.

At last, we mention that strong stochastic properties may be obtained for the asymptotic measure of a hyperbolic attractor. For example, the correlation function may have exponential decay and the central limit theorem may be proved, see Pesin and Sinai (1981). An important tool is using the simple representation in terms of symbolic dynamic and Markov partitions, see also Guckenheimer and Holmes(1990).

In next two sections, we discuss the other two important class of ergodic quantities: the fractal dimensions and entropy.

2.7 Fractal Dimension

Even though real objects are often well approximated by smooth sets, it is argued that many real shapes are actually fractals (see e.g. Mandelbrot 1982, Barnsley 1988). Fractal geometry deals with complicated and irregular sets. For example Cantor sets serve as a pathology in real analysis but are the very basic and prototype of a fractal. There does not appear to exist a universal definition of fractals. In general it has nonintegral dimension (defined below). Usually a fractal set has a very fine structure and has some kind of (e.g. statistical) self-similar property, i.e. the structure of any small portion resemble the whole in some sense(e.g. statistically). Because of the self-similar property, fractals can be easily generated by simple algorithms, for example using the *iterated function systems* of Barnsley(1988).

The main tool of studying fractals is the dimension in its many forms. There are several fractal dimensions, which address different geometrical or statistical aspects of a set. Fractal dimensions can be classified into two categories: one is set(metric)-based dimension such as Hausdorff dimension, the other is measure-based dimension, such as information dimension, pointwise dimension, correlation dimension, and Lyapunov dimension. Generally, it may be expected that the fractal dimensions in each category assume the same value, see Farmer et al (1983), Eckmann and Ruelle (1985), Ruelle (1989). Other references of fractals and fractal dimensions include Mandelbrot (1982), Barnsley (1988), and Falconer (1990).

The importance of fractals in the study of a dynamical system is that, most strange attractors are fractals, e.g. the Hénon, Ikeda, Rössler, Lorenz attractors are all fractals, so that a fractal dimension is often an indication of chaos.

Hausdorff dimension is defined in terms of a metric of a set. Let A be a nonempty set with a metric, and $r > 0$, and denote by $\sigma(r)$ a covering of A by a countable family of sets σ_k with diameter $d_k = \text{diam}\sigma_k \leq r$. Given $\alpha \geq 0$, let

$$m_r^\alpha(A) = \inf_{\sigma(r)} \sum_k (d_k)^\alpha.$$

When $r \downarrow 0, m_r^\alpha(A)$ increases to a limit $m^\alpha(A)$ (possibly infinite) called the Hausdorff measure of A in dimension α. We define

$$\dim_H A = \sup\{\alpha : m^\alpha(A) > 0\}$$

and call this quantity the *Hausdorff dimension of A*. Note that $m^\alpha(A) = +\infty$ for $\alpha < \dim_H A$, and $m^\alpha(A) = 0$ for $\alpha > \dim_H A$.

Given a probability measure ρ, its *information dimension* $\dim_H \rho$ is the smallest Hausdorff dimension of a set S of ρ measure 1. The Hausdorff dimension $\dim_H(\text{supp}\rho)$ of the support of ρ may be strictly larger than $\dim_H \rho$, see p. 641 of Eckmann and Ruelle (1985).

A simple fractal dimension for ρ is given in terms of the probability mass of a small ball. Let $B_r(\mathbf{x})$ be a ball of radius r centered at $\mathbf{x}$, i.e.

$$B_r(\mathbf{x}) = \{\mathbf{y} : d(\mathbf{y}, \mathbf{x}) \leq r\}$$

where d is some distance in $\mathcal{R}^p$ such as the Euclidean distance or the maximum norm. If

$$d = \lim_{r \to 0} \frac{\log \rho(B_r(\mathbf{x}))}{\log r}, \tag{2.9}$$

for ρ-almost all $\mathbf{x}$, and d is independent of $\mathbf{x}$, then d is called the *pointwise fractal dimension* of ρ. In particular, if ρ is ergodic, the existence of limit (2.9) implies that d is independent of $\mathbf{x}$ (see Ott 1993).

When an attractor has the pointwise dimension, it is called a homogeneous fractal. However, it may often happen that

$$\rho[B_{\mathbf{x}}(r)] \sim r^{\alpha(\mathbf{x})},$$

namely the scaling index α depends on $\mathbf{x}$, in which case we call the attractor an *inhomogeneous fractal* or *multifractal*. A multifractal measure ρ is not ergodic. See Ruelle (1989), Ott (1993), Falconer (1990) for more discussions on multifractal measures.

Closely related to pointwise dimension is the correlation dimension. Since $\rho[B_{\mathbf{x}}(r)]$ will in general depend on $\mathbf{x}$, we consider the expectation wrt ρ, i.e.

$$E\rho[B_X(r)] = \int \rho[B_{\mathbf{x}}(r)]\rho(d\mathbf{x}), \tag{2.10}$$

and define the *correlation dimension* as

$$d_c = \lim_{r \to 0} \frac{\log E\rho[B_X(r)]}{\log r}.$$

Another interpretation of (2.10) is given by

$$E1_{\{d(X,Y)\leq r\}} = \int\int 1_{\{d(\mathbf{x},\mathbf{y})\leq r\}}\rho(d\mathbf{x})\rho(d\mathbf{y}),$$

where X, Y are iid with distribution ρ. Above quantity is often called the *correlation integral*, and is easy to estimate from time series. The correlation dimension has perhaps received most attention due to a popular algorithm of Grassberger and Procaccia (1983), which enables us to compute it easily based on time series, see e.g. Eckmann and Ruelle (1985), and Smith (1992a).

Denote the Lyapunov exponents for a map M in $\mathcal{R}^p$ wrt an invariant measure ρ by $\lambda_1 \geq \cdots \geq \lambda_p$ (also collectively called the Lyapunov spectrum). The Lyapunov spectrum not only refers to the dynamical properties of the attractor, it also reveals its geometrical aspects through the Lyapunov dimension.

Denote the sum of the k largest Lyapunov exponents by

$$c_\rho(k) = \sum_{i=1}^{k} \lambda_i.$$

Notice that the maximum of $c_\rho(k)$ is the sum of the positive Lyapunov exponents, and that $c_\rho(k)$ becomes negative for sufficiently large k. (This is the case for dissipative systems, where $c_\rho(p) < 0$ in a $p-$dimensional system.) When $c_\rho(k) \geq 0$ and $c_\rho(k+1) < 0$, we define the *Lyapunov dimension* as

$$\dim_\Lambda \rho = k + \frac{c_\rho(k)}{|\lambda_{k+1}|}.$$

The connection of Lyapunov dimension with other fractal dimensions are discussed in Farmer et al (1983), Eckmann and Ruelle (1985), and Ruelle (1989). In particular, a well-known Kaplan and Yorke conjecture says that the information dimension equals the Lyapunov dimension if ρ is an SRB measure. In some special cases, this conjecture is proved. Refer to Eckmann and Ruelle (1985) for more discussions.

2.8 Entropy

Lyapunov exponents quantify how much sensitivity to initial conditions, or chaos, is present in a system. The entropy, defined as the average rate that information is produced, is closely related to sensitive dependence on initial conditions. Two close values which are indistinguishable at certain precision will become quite distinct at some time later since they diverge exponentially fast in a chaotic system. In this sense information is produced in a chaotic system. For example consider the dynamical system given by $m(x) = 2x (\text{mod } 1)$ for $x \in (0, 1)$. This map has sensitivity to initial conditions, and $\lambda = \log 2$. Clearly any two values which are different but cannot be observed in given precision will become observable at some later time.

If ρ is an ergodic probability measure for a dynamical system M, the concept of mean rate of creation of information $h(\rho)$, often called the *Kolmogorov-Sinai entropy* can be defined for ρ which has a compact set with a given metric. Given any finite partition $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_\alpha)$ of the phase space, define

$$H(\mathcal{A}) = -\sum_{i=1}^{\alpha} \rho(\mathcal{A}_i) \log \rho(\mathcal{A}_i),$$

with the convention that $u \log u = 0$ when $u = 0$. We use $\mathcal{A}^{(n)}$ to denote the partition generated by $\mathcal{A}$ in a time interval of length n through the evolution of M, which consists of all the pieces given by

$$\mathcal{A}_{i_1} \cap M^{-1}\mathcal{A}_{i_2} \cap \cdots \cap M^{-n+1}\mathcal{A}_{i_n}.$$

The following limits can be shown to exist: $h(\rho, \mathcal{A}) = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{A}^{(n)})$, and $h(\rho)$ is defined as the limit of $h(\rho, \mathcal{A})$ as the diameter of the partition $\mathcal{A}$ tends to zero. Clearly, $h(\rho, \mathcal{A})$ is the rate of information creation with respect to the partition $\mathcal{A}$, and $h(\rho)$ its limits for finer and finer partitions. The last limit may be avoided, that is, we actually have $h(\rho, \mathcal{A}) = h(\rho)$. This is the case if $\mathcal{A}$ is a generating partition. This holds in particular if the diameter of $\mathcal{A}^{(n)}$ tends to 0 as $n \to \infty$. Refer to Eckmann and Ruelle (1985) for more details.

The relation of entropy to Lyapunov exponents is very interesting. We have the following general inequality. Let M be a differentiable map of a finite-dimensional manifold and ρ an ergodic measure with compact support. Then

$$h(\rho) \leq \sum_{\lambda_i > 0} \lambda_i. \tag{2.11}$$

The equality corresponding to (2.11) seems to hold often for the physical measures in which we are mainly interested. This equality is called the *Pesin identity*:

$$h(\rho) = \sum_{\lambda_i > 0} \lambda_i. \tag{2.12}$$

An important result due to Pesin is the following: if ρ is invariant under diffeomorphism f, and ρ has smooth density with respect to Lebesgue measure, the Pesin identity holds. More generally, the Pesin identity holds for the SRB measures. The following result due to Ledrappier and Young (1984) is given in Eckmann and Ruelle (1985).

Theorem 2.4 *Let M be a twice differentiable diffeomorphism of an $m-$dimensional manifold D and ρ an ergodic measure with compact support. The following conditions are then equivalent:*

(a) The measure ρ is an SRB measure, i.e. ρ is absolutely continuous along unstable manifolds.

(b) The measure ρ satisfies Pesin's Identity.

Furthermore, if these conditions are satisfied, the density functions defined along unstable manifolds are differentiable.

The estimation of entropy provides another way of quantifying chaos. We refer to Eckmann and Ruelle (1985) for related discussions and references therein.

2.9 State Space Reconstruction

Takens theorem. The state space reconstruction technique used in recovering properties of the original process from an observed scalar time series is justified by Takens' theorem in the deterministic case. Main references include Takens (1981) and Sauer et al (1991).

First we review some basic concepts in differential topology. A smooth map M on D is an *immersion* if the derivative map $D_M(\mathbf{x})$ (given by the Jacobian matrix of M at $\mathbf{x}$) is one-to-one at every point $\mathbf{x}$ of D, or equivalently $D_M(\mathbf{x})$ has full rank on the tangent space. This can happen whether or not M is one-to-one. An *embedding* of D is a smooth diffeomorphism from D to its image $M(D)$, that is, a smooth one-to-one map which has a smooth inverse. The map M is an embedding if and only if it is a one-to-one immersion. We will use the word "*generic*" to mean that there exists an open and dense set in the C^1−topology of maps whose elements satisfy the stated property. We state Takens' theorem (Takens, 1981) in the discrete-time case.

Theorem 2.5 (Takens,1981) *Let D be a compact manifold of dimension d. For pairs $(S, h), S : D \rightarrow D$ a smooth diffeomorphism and $h : D \rightarrow R$ a smooth measurement function, it is a generic property that the delay coordinate map $\Phi(S, h) : D \rightarrow \mathcal{R}^{2d+1}$ given by*

$$\Phi(S, h)\mathbf{x} = (h(\mathbf{x}), h(S\mathbf{x}), h(S^2\mathbf{x}), \ldots, h(S^{2d}\mathbf{x}))$$

is an embedding.

Takens' Theorem is extended to more general cases by Sauer et al (1991), where D is a fractal subset with box-counting dimension d_b, "the generic property" is replaced by "probability-one" in a properly prescribed probability space, and the delay coordinate map $\Phi(S, h) : D \rightarrow R^{p_e}$ is given by

$$\Phi(S, h)\mathbf{x} = (h(\mathbf{x}), h(S\mathbf{x}), h(S^2\mathbf{x}), \ldots, h(S^{p_e-1}\mathbf{x}))$$

where the integral embedding dimension satisfies $p_e > 2d_b$. We refer to Takens (1981), Sauer et al (1991) for more details, including the related theorems for the continuous-time case. In the following we shall discuss the implications of Takens' theorem for data analysis, including estimations of fractal dimension, entropy, and Lyapunov exponents.

The following diagram shows the process of embedding the observation process $\{h(s^n x)\}$ into a higher-dimensional process in $\mathcal{R}^{p_e}$

$$\begin{tikzcd}
D \arrow[r, "\Phi(S,h)"] \arrow[d, "S"'] & \Phi(D) \arrow[r, phantom, "\subset"] \arrow[d, "M"] & \mathcal{R}^{p_e} \\
D \arrow[r, "\Phi(S,h)"'] & \Phi(D) \arrow[r, phantom, "\subset"] & \mathcal{R}^{p_e}
\end{tikzcd}$$

where M is the induced map on the embedded space $\mathcal{R}^{p_e}$ defined by $M = \Phi S \Phi^{-1}$. The existence of Φ^{-1} or M is ensured if $p_e \geq 2d + 1$ from Takens' theorem. We have the following relation among the maps

$$M\Phi = \Phi S, \tag{2.13}$$

and in general we have

$$M^n \Phi = \Phi S^n. \tag{2.14}$$

Let's look closely at how the reconstructed map M can given in the embedded space $\mathcal{R}^{p_e}$. From (2.13), we have

$$M(\Phi(S,h)\mathbf{x}) = \Phi(S,h)(S\mathbf{x}) = (h(S\mathbf{x}), h(S^2\mathbf{x}), \ldots, h(S^{p_e}\mathbf{x}))$$

Clearly, M has the form

$$M(x_0, x_1, \ldots, x_{p_e-1}) = (x_1, x_2, \ldots, m(x_0, x_1, \ldots, x_{p_e-1})), \tag{2.15}$$

for any point $(x_0, x_1, \ldots, x_{p_e-1})$ in $\mathcal{R}^{p_e}$, where $m : \mathcal{R}^{p_e} \to \mathcal{R}$ is a smooth function given by the univariate time series

$$x_t = m(x_{t-p_e}, x_{t-p_e+1}, x_{t-1}),\ t = 0, 1, 2, \ldots. \tag{2.16}$$

It is noted that, the fact that Takens' reconstruction map M has the simple form of (2.15) is from the use of the delay coordinate technique. This may be another advantage for use of the delay coordinate method in reconstructing the underlying dynamics. As a result, the difficulty of reconstructing a p_e-dimensional map is reduced to the problem of estimating only a p_e-variate real function m. The latter is a problem of multiple regression estimation.

Recovery of dynamical properties in time series. We now show how the geometric and dynamical properties including fractal dimensions, entropy, and Lyapunov exponents of the original system f can be derived from that of the reconstructed delay map m in (2.16) or M in (2.15). These facts are often taken for granted in the literature. However, they do not follow immediately from Takens' theorem. The needed arguments will be stated below, but first some more concepts in ergodic theory are needed.

Let $S : X \to X$ and $M : Y \to Y$ be two maps. Recall we say that S and M are topologically conjugate if there exists a homeomorphism $\Phi : X \to Y$ such that, $\Phi S = M\Phi$. The homeomorphism Φ is called a topological conjugacy. Mappings which are topologically conjugate are completely equivalent in terms of their dynamics. For example, if S is topologically conjugate to G via Φ, and x is a fixed point for S, then Φx is fixed for M. Indeed, $\Phi x = \Phi(Sx) = \Phi Sx = M\Phi x$. Similarly, Φ gives a one-to-one correspondence between periodic points of S and periodic points of M. One may also also check that eventually periodic and asymptotic orbits for S go over via Φ to similar orbits for M, and that S is topologically transitive(i.e. S has dense orbits) if and only if M is. For more discussions of topological conjugacy, and its important applications in the analysis of dynamics via symbolic dynamics theory, we refer to the excellent book by Devaney (1989). Next we will discuss the measure-theoretic consequences of topological conjugacy.

Let $(X, \mathcal{A}, \mu)$ be a probability space, let $M : X \to X$ be a one-to-one onto map such that both M and M^{-1} are measurable: $M^{-1}\mathcal{A} = M\mathcal{A} = \mathcal{A}$. Assume further that $\mu(M^{-1}E) = \mu(E)$ for all $E \in \mathcal{A}$, that is, M is a measure-preserving transformation. This system will be denoted by $(X, \mathcal{A}, \mu, M)$. We say that two systems $(X, \mathcal{A}, \mu, S)$ and $(Y, \mathcal{B}, \nu, M)$ are *metrically isomorphic* if there is a one-to-one onto map $\Phi : X \to$

Y such that both Φ and Φ^{-1} are measurable,

$$\Phi S = M\Phi \text{ on } X,$$

and

$$\mu(\Phi^{-1}E) = \nu(E) \text{ for all } E \in \mathcal{B}. \tag{2.17}$$

The map Φ is called an *isomorphism.*

Isomorphic systems share many ergodic properties. An important isomorphism invariant the entropy $h(M)$ of M defined in Section 2.8. The entropy can be used to distinguish some nonisomorphic systems, and is in fact a complete isomorphism invariant within certain classes of systems. The measure-theoretic fractal dimensions and entropy for the asymptotic measure of S will be the same as that given by M, if the corresponding measures are isomorphic. The invariant measures for S and M are isomorphic, if they are defined by time averages of corresponding orbits through Φ. Next, we shall discuss how the Lyapunov spectrum can preserved, where the differential structure will be involved.

Since they are smooth, the tangent maps given by the Jacobian matrices for M, Φ, S satisfy

$$D_M(\Phi\mathbf{x})D_\Phi(\mathbf{x}) = D_\Phi(S\mathbf{x})D_S(\mathbf{x}), \tag{2.18}$$

for every $\mathbf{x} \in D$ from (2.13). It is noted that $D_\Phi(\mathbf{x})$ has full rank but may not be a square matrix, so its inverse may not be defined. More generally, we have the following which corresponds to (2.14),

$$D_M(\Phi S^{n-1}\mathbf{x})\cdots D_M(\Phi S\mathbf{x})D_M(\Phi\mathbf{x})D_\Phi(\mathbf{x}) = D_\Phi(S^n\mathbf{x})D_S(S^{n-1}\mathbf{x})\cdots D_S(S\mathbf{x})D_S(\mathbf{x}). \tag{2.19}$$

From (2.18) and (2.19), it is seen that if $\mathbf{x}$ is a fixed point (or a periodic point with period n), and if D_Φ is a square matrix and hence has an inverse, since $D_M(\Phi\mathbf{x})$ is similar to $D_S(\mathbf{x})$ (or $D_{M^n}(\Phi\mathbf{x})$ is similar to $D_{S^n}(\mathbf{x})$), their eigenvalues will remain the same, and consequently, the Lyapunov exponents at $\mathbf{x}$, defined by the logarithms of modulus of the eigenvalues, are preserved. More generally, for any given point $\mathbf{x}$, if

S has Lyapunov spectrum $\lambda_1(\mathbf{x}) \geq \lambda_2(\mathbf{x}) \geq \ldots \geq \lambda_p(\mathbf{x})$ at $\mathbf{x}$, the first p Lyapunov exponents of M from the corresponding orbit through Φ will be the same. Under an ergodic measure ρ, S has Lyapunov spectrum which is independent of the initial values, the first p Lyapunov exponents of M will be the same as those of S under the corresponding ergodic measure $\rho\Phi^{-1}$.

Chapter 3

MULTIVARIATE LOCALLY WEIGHTED REGRESSION

3.1 Introduction

Regression models have been used in modeling dependence relationships among variables. Suppose we are given data $(Y_1, X_1), \ldots, (Y_n, X_n)$ which are assumed to be a set of independent and identically distributed $\mathcal{R}^{p+1}$-valued random vectors, where the Y_i's are the scalar response variables and the X_i's are the $\mathcal{R}^p$-valued predictor variables with density function f. The multivariate (mean) regression problem is that of estimating the conditional mean function

$$m(\mathbf{x}) = E(Y|X = \mathbf{x})$$

at a $p-$dimensional point $\mathbf{x}$ in the support of f. We denote the conditional variance function by $\nu(\mathbf{x}) = \mathrm{Var}(Y|X = \mathbf{x})$. A specific model formulation is given by:

$$Y_i = m(X_i) + \nu^{1/2}(X_i)\varepsilon_i, \quad i = 1, \ldots, n \tag{3.1}$$

where ε_i's are independent and identically distributed scalar random variables with $E(\varepsilon_i|X_i) = 0$ and $\mathrm{Var}(\varepsilon_i|X_i) = 1$. The above setup with X_i's being random variables is called the *random design* model. When $\{\mathbf{x}_i\}$ are predetermined quantities, and $\{Y_i\}$ are the corresponding response random variables, a model similar to (3.1) is formulated as

$$Y_i = m(\mathbf{x}_i) + \nu^{1/2}(\mathbf{x}_i)\varepsilon_i, \quad i = 1, \ldots, n,$$

where $m(\mathbf{x})$ is the mean of Y for any given $\mathbf{x}$, and $\{\varepsilon_i\}$ are iid with zero mean and unit variance. This latter formulation is called the *fixed design* model. Since we will be concerned with applications in time series which belong to the random design setup, we will only state results for the random design model. However, the methods presented here are equally applicable to the fixed design model as well.

Since in many applications a parametric form is not known, considerable interest has centered on nonparametric regression. In a nonparametric setup, a parametric functional form is not specified, instead m is only assumed to satisfy some general smoothness conditions. The mostly studied nonparametric regression methods include the kernel method and the smoothing spline, see Härdle (1990), Müller (1988), and Rosenblatt (1991).

The kernel estimators seem to be the simplest and are often used. There are two popular kernel estimators of regression, one often called the *Nadaraya-Watson estimator* $\hat{m}_{NW}$, the other the *Gasser-Müller estimator* $\hat{m}_{GM}$. For simplicity, we only discuss the two estimators with univariate predictor (p=1). The Nadaraya-Watson estimator $\hat{m}_{NW}$ is given by

$$\hat{m}_{NW}(x) = \frac{(nh)^{-1}\sum_{i=1}^{n} K((X_i - x)/h)Y_i}{(nh)^{-1}\sum_{i=1}^{n} K((X_i - x)/h)},$$

where K is a kernel function, often given by a density function, and h is the bandwidth parameter, which controls the neighborhood of data points used. $\hat{m}_{NW}$ generalizes the concept of local averages or means. Another interpretation is that $\hat{m}_{NW}$ minimizes the weighted sum of squares $\sum_{i=1}^{n}(Y_i - a)^2 K(\frac{X_i - x}{h})$. The Gasser-Müller estimator $\hat{m}_{GM}$ is given by

$$\hat{m}_{GM}(\mathbf{x}) = \sum_{i=1}^{n} Y_{(i)} \int_{t_{i-1}}^{t_i} h^{-1}K(x - t/h)dt,$$

where $t_0 = -\infty, t_n = \infty, t_i = (X_{(i)} + X_{(i+1)})/2$ for $i = 1, \ldots, n-1$, and $X_{(1)}, \ldots, X_{(n)}$ are order statistics among the X's and $Y_{(1)}, \ldots, Y_{(n)}$ are the corresponding Y's.

Chu and Marron (1992) have compared the two regression smoothers based on the criterion of (pointwise) asymptotic mean squared error, which comprise two parts, the squared bias and the variance. Under some regularity conditions, for an interior point x, the bias of the two estimators have the following expansions corresponding to (4.3) and (4.4) of Chu and Marron (1992).

$$\begin{aligned}\text{Bias}(\hat{m}_{NW}(x)) &= \frac{1}{2}\mu_2 h^2 \{m^{(2)}(x) + 2m'(x)\frac{f'(x)}{f(x)}\} + O(n^{-1/2}h^{1/2}) + o(h^2), \\ \text{Bias}(\hat{m}_{GM}(x)) &= \frac{1}{2}\mu_2 h^2 m^{(2)}(x) + O(n^{-1}) + o(h^2),\end{aligned}$$

where $m'(x), m^{(2)}(x)$ denote first and second derivatives of m at x, etc., and $\mu_2 = \int u^2 K(u)du$. It is easy to see that the biases are not comparable. In particular, the asymptotic bias of $\hat{m}_{NW}$ involves the extra term $\mu_2 h^2 m'(f'/f)$, which be undesirable in some sense, e.g. in a minimax-theoretic sense as demonstrated by Fan (1993). On the other hand, the variances of the two estimators have expansions which correspond to (3.5) and (3.6) in Chu and Marron (1992):

$$\text{Var}(\hat{m}_{NW}) = \frac{\nu(x)J_0}{nhf(x)} + o((nh)^{-1}),$$

$$\text{Var}(\hat{m}_{GM}) = \frac{3}{2}\frac{\nu(x)J_0}{nhf(x)} + o((nh)^{-1}),$$

where $J_0 = \int K(u)^2 du$. It is seen that $\hat{m}_{GM}$ is not as efficient as $\hat{m}_{NW}$ in terms of the asymptotic variance.

A resolution to the choice of kernel estimators is given by Fan (1993), who has made the interesting discovery that the *local linear regression estimator*, which will be defined in Section 3.3, can have advantages of both estimators, namely the local linear smoother attains the same asymptotic bias expression of $\hat{m}_{GM}$ while maintaining the same asymptotic variance as $\hat{m}_{NW}$. See Section 3.3 for more details. Fan also finds that the local linear regression estimator has some nice optimality properties in terms of minimax efficiency, see Fan (1993) for more details. Moreover, the nice properties of the local linear regression estimator at an interior point also carry over to a boundary point. In particular, the local linear estimator automatically has bias of order $O(h^2)$ near the boundary, so it is free from the boundary effect which happens with $\hat{m}_{NW}, \hat{m}_{GM}$. See Fan and Gijbels (1992) for further details.

The local linear regression estimator, given by the local linear fit, is a special case of the locally weighted regression estimators or estimators from the locally weighted polynomial fit, as studied in Stone (1977, 1980), Cleveland (1979), Cleveland and Devlin (1988), and Ruppert and Wand (1994). The local polynomial method, particularly the local linear fit, is easy to implement. This is particularly important in the multivariate case, where computation is often one of the main concerns. Multivariate locally weighted regression is studied in Cleveland and Devlin (1988), and Ruppert and Wand (1994), where the local linear fit and the local quadratic fit are considered.

Nonparametric derivative estimation of a regression function is often required for constructing confidence intervals or bands for a regression function or in a bandwidth selection procedure for a nonparametric regression. It also has some real applications, for example in the study of growth curves in a longitudinal study (e.g. Müller 1988). Our interest is motivated by application in estimating the Lyapunov exponents, which are defined in terms of the partial derivatives of the autoregression function, see Chapter 5.

Nonparametric regression estimates of derivatives have often been studied in a fixed design context, e.g. Müller (1988). It is usually given by either differentiating regression estimators, or using a higher-order kernel. A drawback with these estimators is that they often have complicated form and are not easy to analyze, particularly in the case of random design setup, which is our main concern. On the other hand, Ruppert and Wand (1994) have studied the local polynomial estimators in the univariate case, which turns out to be easy to interpret and to study. Moreover, the boundary effect in connection with traditional nonparametric derivative estimation is eliminated.

In this chapter, the locally weighted polynomial fit is studied in the multivariate predictor case. Two important cases, the local linear fit and local quadratic fit, are studied in detail. We will derive the large-sample behavior of the conditional bias and conditional covariance matrix of the regression and derivative estimators for both cases. The results on regression estimators have appeared in more general form in Ruppert and Wand. The problem on partial derivative estimation seems to be considered for the first time here, and our results generalize directly Ruppert and Wand's results in the univariate case. The results will also be useful for the bandwidth selection problem. Results of this chapter will be generalized to dependent data in Chapter 4.

This chapter is structured as follows. Section 3.2 gives some notations and preliminaries. Section 3.3 discusses the local linear fit. Section 3.4 discusses the local quadratic fit. The proofs of theorems are given in Section 3.6.

3.2 Some Notations

Given a $p \times p$ matrix A, A^T denotes its transpose. For $A_1, \cdots, A_p (p > 1)$ which are square matrices or numbers, we denote

$$\mathrm{diag}\{A_1, \cdots, A_p\} = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_p \end{pmatrix},$$

where the suppressed elements are zeros.

We denote vec$A$ as the column vector stacking the columns of A. That is, let $A = (a_1, \cdots, a_q)$, then

$$\mathrm{vec}A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{pmatrix}.$$

If $A = (a_{ij})$ is a symmetric matrix of order p, we use vech$A$ to denote the column vector of dimension $p(p+1)/2$ by stacking the elements on and below the diagonal column by column, that is

$$\mathrm{vech}A = (a_{11}, \cdots, a_{p1}, a_{22}, \cdots, a_{p2}, \cdots, a_{pp})^T.$$

$\mathrm{vech}^T A$ is used as a shorthand notation for $\{\mathrm{vech}A\}^T$.

We need some notations and preliminaries in multivariate calculus. For a given point $\mathbf{x} = (x_1, \cdots, x_p)^T$ in $\mathcal{R}^p$, let U denote an open neighborhood of $\mathbf{x}$, let $C^d(U)$ denote the class of functions whose (mixed) partial derivatives up to dth order exist and are continuous in U. For $g(x_1, \cdots, x_p) \in C^d(U)$, define the first-order differential operator $\frac{\partial}{\partial x_i}$'s by

$$\frac{\partial}{\partial x_i} g(\mathbf{x}) = \frac{\partial g(\mathbf{x})}{\partial x_i}, 1 \le i \le p,$$

define product of two differential operators D_1, D_2 by

$$D_1 D_2 g(\mathbf{x}) = D_1(D_2 g(\mathbf{x})), \text{ for any } g \in C^d(U),$$

where D_1, D_2 are any of $\frac{\partial}{\partial x_i}$'s. Then higher-order differentials are defined based on products of first-order differentials, e.g. the second-order partial derivative operators are given by

$$\frac{\partial^2}{\partial x_i^2} = (\frac{\partial}{\partial x_i})^2, \frac{\partial^2}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}, i \ne j,$$

and more generally denote the lth-order partial derivative operators by

$$\frac{\partial^l}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_p^{i_p}} = \frac{\partial}{\partial x_1^{i_1}}(\frac{\partial^{l-1}}{\partial x_2^{i_2} \cdots \partial x_p^{i_p}}),$$

where $i_1, \ldots, i_p$ are nonnegative integers and $i_1 + \cdots + i_d = l$. The lth-order operator on function g is defined by

$$(\frac{\partial^l}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_p^{i_p}}) g(\mathbf{x}) = \frac{\partial^l g(\mathbf{x})}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_p^{i_p}},$$

for any $g \in C^{(d)}(U), l \le d$. The lth-order differential of g at $\mathbf{x}$ is given by

$$(\sum_{i=1}^{p} \frac{\partial}{\partial x_i} u_i)^l g(\mathbf{x}) = \sum_{i_1, \ldots, i_p} C^l_{i_1 \cdots i_p} \frac{\partial^l g(\mathbf{x})}{\partial x_1^{i_1} \partial x_2^{i_2} \cdots \partial x_p^{i_p}} (u_1)^{i_1} \cdots (u_p)^{i_p},$$

where the summations are over all distinct nonnegative integers $i_1, \ldots, i_p$ such that $i_1 + \cdots + i_p = l$, and $C^l_{i_1 \cdots i_p} = l!/(i_1! \cdots i_p!)$.

Given a random sequence $\{a_n\}$, we denote $a_n = o_p(\gamma_n)$ if $\gamma_n^{-1} a_n$ tends to zero in probability. We denote $a_n = O_p(\gamma_n)$ if $\gamma_n^{-1} a_n$ is bounded in probability, that is, for any $\epsilon > 0$, there exists an M_ϵ, N_ϵ such that

$$P(|\gamma_n^{-1} a_n| > M_\epsilon) < \epsilon, \text{ for all } n > N_\epsilon.$$

The concept of stochastic order can be extended to a sequence of vectors or matrices $\{A_n\}$, through definition of its components. In particular, we denote a sequence of $p \times p$ matrices $A_n = O_p(\gamma_n)$ if and only if each component $A_n(i,j) = O_p(\gamma_n), i, j = 1, \ldots, p$. Letting $\|A\|$ denote a norm of a matrix A, say $\|A\| = (\sum_{i,j=1}^p A(i,j)^2)^{1/2}$, then $A_n = O_p(\gamma_n)$ if and only if $\|A_n\| = O_p(\gamma_n)$.

3.3 Local Linear Fit

In this section we study the local linear fit with multivariate predictor variables. Specifically, at any given $\mathbf{x}$, the local linear estimators of regression and partial derivatives are derived by minimizing the weighted sum of squares

$$\sum_{i=1}^{n} \{Y_i - a - b^T(X_i - \mathbf{x})\}^2 K(\frac{X_i - \mathbf{x}}{h}). \tag{3.2}$$

where $K(\cdot)$ is the weighting function, h is the bandwidth, and a and b are parameters. The solution $\hat{a} = \widehat{m(x)}$ is an estimate of $m(\mathbf{x})$, and $\hat{b} = \widehat{D_m(\mathbf{x})}$ is an estimate of $D_m(\mathbf{x}) = (\partial m(\mathbf{x})/\partial x_1, \cdots, \partial m(\mathbf{x})/\partial x_p)^T$. Written in matrix form, the local linear estimator $\hat{\beta} = (\hat{a}, \hat{b}^T)^T$ is given by:

$$\hat{\beta} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W Y$$

where $Y = (Y_1, \cdots, Y_n)^T, W = \mathrm{diag}\{K(\frac{X_1 - \mathbf{x}}{h}), \cdots, K(\frac{X_n - \mathbf{x}}{h})\}$, and we use $\mathbf{X}$ to denote the $n \times (p+1)$ design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (X_n - \mathbf{x})^T \end{pmatrix}.$$

The weighting or kernel function K is generally a density function. For example one possible choice is the standard multivariate normal $K_g(\mathbf{x}) = (2\pi)^{-p/2} \exp(-\|\mathbf{x}\|^2/2)$. The uniform kernel $K_u(\mathbf{x}) = C_b^{-1} 1_{\{\|\mathbf{x}\| \le 1\}}$ is often used by practical workers, where $C_b = \pi^{p/2}/\Gamma(\frac{p+2}{2})$ is the volume of the p-dimensional unit sphere.

A large family of kernels is given by

$$K_{\alpha\beta}(\mathbf{x}) = \begin{cases} C_{\alpha\beta}^{-1}(1 - \|\mathbf{x}\|^\alpha)^\beta, & \text{if } \|\mathbf{x}\| \le 1, \\ 0, & \text{otherwise}, \end{cases}$$

for $\beta > -1, \alpha > 0$, where $C_{\alpha\beta}$ is the normalizing constant which makes K integrate to one, and is given by: $C_{\alpha\beta} = 2\pi^{\frac{p}{2}} B(\beta + 1, \frac{p}{\alpha})/(\alpha \Gamma(\frac{p}{2}))$, where B is the beta function, Γ is the gamma function.

Some important cases:

$$\begin{aligned} &K_{21}(\alpha = 2, \beta = 1): && \text{Epanechnikov;} \\ &K_{22}(\alpha = 2, \beta = 2): && \text{biweight;} \\ &K_{23}(\alpha = 2, \beta = 3): && \text{triweight;} \\ &K_{33}(\alpha = 3, \beta = 3): && \text{tricube.} \end{aligned}$$

In practice, the choice of a particular kernel is often based on considerations of smoothness and computational efficiency. For example, with some care in programming, kernels of finite support speed up computation considerably.

Instead of using $h^{-p} K((X - \mathbf{x})/h)$, more general $|B|^{-1} K(B^{-1}(X - \mathbf{x}))$, where B is a positive definite matrix, can be used as our weighting function. We will not pursue this generalization further, but refer interested readers to Ruppert and Wand (1994) for related discussions.

For simplicity, from now on K will be assumed to be spherically symmetric, i.e. there exists a function k such that $K(\mathbf{x}) = k(\|\mathbf{x}\|)$, where $\|\cdot\|$ is the Euclidean norm. Without affecting the estimators, K is normalized to integrate to 1. Most kernels in use, including all the kernels given above, satisfy these conditions.

Since the unconditional moments of the estimators considered here may not exist in general, in this chapter we will work with their conditional moments (given observations of predictor variables X_i's). More insights are given by considering the large-sample behavior as $h \to 0, nh^p \to \infty$. The following theorem gives the large-sample expansions for the conditional bias and conditional covariance matrix of the local linear estimators. We assume that the kernel K is a spherically symmetric density function and satisfies the moment condition,

$$\int u_1^8 K(u_1, \cdots, u_p) du_1 \cdots du_p < \infty. \tag{3.3}$$

Theorem 3.1 *For an interior point* $\mathbf{x} = (x_1, \cdots, x_p)^T$ *inside the support of design density* f *and* $f(\mathbf{x}) > 0, \nu(\mathbf{x}) > 0$*, if there exists an open neighborhood* U *of* $\mathbf{x}$ *such that* $m \in C^3(U), f \in C^1(U), \nu \in C^0(U)$ *then for* $h \to 0, nh^p \to \infty$ as $n \to \infty$*, the conditional bias of the local linear regression and partial derivative estimators have the asymptotic expansions*

$$E\left\{ \left(\begin{array}{c} \widehat{m(\mathbf{x})} \\ \widehat{D_m(\mathbf{x})} \end{array} \right) - \left(\begin{array}{c} m(\mathbf{x}) \\ D_m(\mathbf{x}) \end{array} \right) \middle| X_1, X_2, \ldots, X_n \right\}$$

$$= \left(\begin{array}{l} \frac{1}{2}h^2\mu_2\nabla_m^2(\mathbf{x}) + h^2(o(h) + O_p(\{nh^p\}^{-\frac{1}{2}})) \\ \frac{h^2}{3!\mu_2}b(m,K) + \frac{h^2}{2\mu_2 f(\mathbf{x})}b_1(m,K) + h(o(h) + O_p(\{nh^p\}^{-\frac{1}{2}})) \end{array} \right), \tag{3.4}$$

where

$$\nabla_m^2(\mathbf{x}) = \sum_{i=1}^p \frac{\partial^2 m(\mathbf{x})}{\partial x_i^2},$$

$$b(m,K) \;=\; \int_{\mathcal{R}^p} \mathbf{u}(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})d\mathbf{u}$$

$$= \begin{pmatrix} \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_1^3} + 3\mu_2^2 \sum_{i=2}^{p} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \partial x_1} \\ \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_2^3} + 3\mu_2^2 \sum_{i \neq 2} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \partial x_2} \\ \vdots \\ \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_p^3} + 3\mu_2^2 \sum_{i=1}^{p-1} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \partial x_p} \end{pmatrix} \tag{3.5}$$

and

$$\begin{aligned} b_1(m,K) &= \int_{\mathcal{R}^p} \mathbf{u}\{\mathbf{u}^T H_m(\mathbf{x})\mathbf{u}\} K(\mathbf{u}) D_f^T(\mathbf{x})\mathbf{u}d\mathbf{u} \\ &\quad - \mu_2\{\int_{\mathcal{R}^p} \mathbf{u}^T H_m(\mathbf{x})\mathbf{u}K(\mathbf{u})d\mathbf{u}\} D_f(\mathbf{x}) \\ &= \begin{pmatrix} (\mu_4 - \mu_2^2)\frac{\partial^2 m(\mathbf{x})}{\partial x_1^2}\frac{\partial f(\mathbf{x})}{\partial x_1} + 2\mu_2^2 \sum_{i=2}^{p} \frac{\partial^2 m(\mathbf{x})}{\partial x_1 \partial x_i}\frac{\partial f(\mathbf{x})}{\partial x_1} \\ (\mu_4 - \mu_2^2)\frac{\partial^2 m(\mathbf{x})}{\partial x_2^2}\frac{\partial f(\mathbf{x})}{\partial x_2} + 2\mu_2^2 \sum_{i \neq 2} \frac{\partial^2 m(\mathbf{x})}{\partial x_2 \partial x_i}\frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ (\mu_4 - \mu_2^2)\frac{\partial^2 m(\mathbf{x})}{\partial x_p^2}\frac{\partial f(\mathbf{x})}{\partial x_p} + 2\mu_2^2 \sum_{i=1}^{p-1} \frac{\partial^2 m(\mathbf{x})}{\partial x_p \partial x_i}\frac{\partial f(\mathbf{x})}{\partial x_p} \end{pmatrix} \end{aligned}$$

The conditional variance-covariance matrix has the asymptotic expansion

$$\begin{aligned} &\text{Cov}\left\{ \begin{pmatrix} \widehat{m(\mathbf{x})} \\ \widehat{D_m(\mathbf{x})} \end{pmatrix} \middle| X_1, X_2, \ldots, X_n \right\} = \\ &\begin{pmatrix} \frac{\nu(\mathbf{x})J_0}{nh^p f(\mathbf{x})} & 0 \\ 0 & \frac{\nu(\mathbf{x})J_2}{\mu_2{}^2 f(\mathbf{x}) nh^{p+2}} I \end{pmatrix} \\ &+\frac{1}{nh^p}\text{diag}\{1, h^{-1}I\}[o(1) + O_p(\{nh^p\}^{-\frac{1}{2}})]\text{diag}\{1, h^{-1}I\}, \end{aligned} \tag{3.6}$$

where I is the identity matrix of dimension p,

$$\mu_\ell = \int u_1^\ell K(u_1, u_2, \cdots, u_p) du_1 du_2 \cdots du_p,$$

$$J_\ell = \int u_1^\ell K^2(u_1, u_2, \cdots, u_p) du_1 du_2 \cdots du_p,$$

for any nonnegative integers ℓ.

REMARKS ON THEOREM 3.1:

1. For the results on the regression estimator to hold only, the assumptions $m \in C^3(U), f \in C^1(U)$ are not necessary. Instead weaker assumptions such as $m \in C^2(U)$, and $f \in C^{(0)}(U)$ will suffice.

2. With this theorem, the asymptotic expansion for the conditional mean squared error (CMSE) of the partial derivative estimator at an interior point can be easily given. The optimal h which will balance the asymptotic conditional bias and asymptotic conditional covariance matrix is given by the rate $n^{-\frac{1}{p+6}}$. The convergence rate in the sense of CMSE of the estimator using the optimal h is seen to be of order $n^{-\frac{2}{p+6}}$, which attains the optimal rate established in Stone (1980).

3. In the special case $p = 1$, the conditional mean squared error of $\widehat{m'(x)}$ is approximated by

$$h^4\{\frac{\mu_4}{3!\mu_2}m^{(3)}(x) + \frac{\mu_4 - \mu_2^2}{2\mu_2}\frac{m^{(2)}(x)f'(x)}{f(x)}\}^2 + \frac{\nu(x)J_2}{{\mu_2}^2 f(x)nh^3}.$$

3.4 Local Quadratic Fit

In this section we consider the main case, the local quadratic fit with multivariate predictors. The local quadratic estimator is derived by minimizing the weighted sum of squares

$$\sum_{i=1}^{n}\{Y_i - a - b^T(X_i - \mathbf{x}) - (X_i - \mathbf{x})^T L(X_i - \mathbf{x})\}^2 K(\frac{X_i - \mathbf{x}}{h}), \tag{3.7}$$

where a is a real number, b is a $p-$dimensional vector, and L is restricted to be a lower triangular matrix for identifiability. Note that the solution $\widehat{a} = \widehat{m(\mathbf{x})}$ is an estimate of regression function, $\hat{b}$ corresponds to an estimate of $D_m(\mathbf{x})$, and $\hat{L}$ corresponds to estimates of elements in the Hessian matrix of $H_m(\mathbf{x})$. Let $H_m(\mathbf{x}) = (h_{ij})$ is the Hessian, $L(x) = (l_{ij})$ satisfies

$$l_{ij} = \begin{cases} h_{ij} & \text{if } i > j \\ h_{ii}/2 & \text{if } i = j \\ 0 & \text{if } i < j \end{cases}$$

Note that the number of parameters in a, b, L is given by $q = \frac{1}{2}(p + 2)(p + 1)$. The following table illustrates values of q, p for $p \leq 10$, which demonstrates the polynomial growth of parameters with respect to the number of predictor variables.

dimension	1	2	3	4	5	6	7	8	9	10
parameters	3	6	10	15	21	28	36	45	55	66

Denote the solution by

$$\widehat{\beta} = (\hat{a}, \hat{b}^T, \mathrm{vech}^T\{\hat{L}\})^T.$$

Then $\hat{\beta}$ is given in matrix form by

$$\hat{\beta} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W Y \tag{3.8}$$

where $Y = (Y_1, \cdots, Y_n)^T, W = \mathrm{diag}\{K(\frac{X_1 - \mathbf{x}}{h}), \cdots, K(\frac{X_n - \mathbf{x}}{h})\}$, and

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - \mathbf{x})^T & \mathrm{vech}^T\{(X_1 - \mathbf{x})(X_1 - \mathbf{x})^T\} \\ \vdots & \vdots & \vdots \\ 1 & (X_n - \mathbf{x})^T & \mathrm{vech}^T\{(X_n - \mathbf{x})(X_n - \mathbf{x})^T\} \end{pmatrix}_{n \times q}.$$

We will assume the kernel K is spherically symmetric as in Section 3.3. We also assume that K satisfies the moment condition,

$$\int u_1^{12} K(u_1, \cdots, u_p) du_1 \cdots du_p < \infty.$$

As in Section 3.3, our purpose is to study the conditional bias and conditional covariance matrix of $\hat{\beta}$. In particular, the conditional bias is given by

$$E(\hat{\beta} \mid X_1, \cdots, X_n) - \beta = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W (M - \mathbf{X}\beta), \tag{3.9}$$

and the conditional covariance matrix is given by

$$\mathrm{Cov}(\hat{\beta} \mid X_1, \cdots, X_n) = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W V W \mathbf{X} (\mathbf{X}^T W \mathbf{X})^{-1}, \tag{3.10}$$

where

$$M = (m(X_1), \cdots, m(X_n))^T, V = \mathrm{diag}\{\nu(X_1), \cdots, \nu(X_n)\}.$$

The large sample asymptotic expansions for the conditional bias and covariance matrix of the local quadratic regression and derivative estimators are given in the following theorem.

Theorem 3.2 *For an interior point $\mathbf{x}$ inside the support of design density f and $f(\mathbf{x}) > 0, \nu(\mathbf{x}) > 0$, if there exists an open neighborhood U of $\mathbf{x}$ such that $m \in C^4(U), f \in C^1(U), \nu \in C^0(U)$, then for $h \to 0, nh^p \to \infty \text{ as } n \to \infty$, the conditional bias of the local quadratic regression and derivative estimators have the asymptotic expansion:*

$$E\left\{\left.\begin{pmatrix} \widehat{m(\mathbf{x})} \\ \widehat{D_m(\mathbf{x})} \\ \mathrm{vech}\{\widehat{L(\mathbf{x})}\} \end{pmatrix} - \begin{pmatrix} m(\mathbf{x}) \\ D_m(\mathbf{x}) \\ \mathrm{vech}\{L(\mathbf{x})\} \end{pmatrix}\right| X_1, X_2, \ldots, X_n \right\} =$$

$$\begin{pmatrix} \frac{h^4}{4!}\theta(m,K) + \frac{h^4}{3!f(\mathbf{x})}\theta_1(m,K) \\ \frac{h^2}{3!\mu_2}b(m,K) \\ h^2\gamma(m,K) + \frac{h^2}{f(\mathbf{x})}\gamma_1(m,K) \end{pmatrix} \tag{3.11}$$

$$+\,\mathrm{diag}\{h^3, h^2 I_1, hI_2\}[o(h) + O_p(\{nh^p\}^{-\frac{1}{2}})],$$

where I_1, I_2 are the identity matrices of dimensions p and $p(p+1)/2$, respectively, $b(m,K)$ is defined in (3.5) in Theorem 3.1 of Section 3.3, and

$$\theta(m,K) = \frac{\mu_4^2 - \mu_2\mu_6}{\mu_4 - \mu_2^2}\sum_{i=1}^{p}\frac{\partial^4 m(\mathbf{x})}{\partial x_i^4} - 6\mu_2^2\sum_{1\le i<j\le p}\frac{\partial^4 m(\mathbf{x})}{\partial x_i^2 \partial x_j^2},$$

$$\theta_1(m,K) = \frac{\mu_4^2 - \mu_2\mu_6}{\mu_4 - \mu_2^2}\sum_{i=1}^{p}\frac{\partial^3 m(\mathbf{x})}{\partial x_i^3}\frac{\partial f(\mathbf{x})}{\partial x_i} - 3\mu_2^2\sum_{\substack{1\le i,j\le p \\ i\ne j}}\frac{\partial^3 m(\mathbf{x})}{\partial x_i \partial x_j^2}\frac{\partial f(\mathbf{x})}{\partial x_i},$$

and $\gamma(m,K)$ and $\gamma_1(m,K)$ are vectors of dimension $p(p+1)/2$ with components

$$\gamma(m,K) = (\gamma_{11}, \cdots, \gamma_{p1}, \gamma_{22}, \cdots, \gamma_{p2}, \cdots, \gamma_{pp})^T,$$

$$\begin{aligned} \gamma_{ii} &= \frac{1}{4!}\frac{\mu_6 - \mu_2\mu_4}{\mu_4 - \mu_2^2}\frac{\partial^4 m(\mathbf{x})}{\partial x_i^4} + \frac{1}{4}\frac{\mu_2\mu_4}{\mu_4 - \mu_2^2}\sum_{\substack{1\le k\le p \\ k\ne i}}\frac{\partial^4 m(\mathbf{x})}{\partial x_i^2 \partial x_k^2}, \\ & \textit{for } 1 \le i \le p. \\ \gamma_{ij} &= \frac{\mu_4}{3!\mu_2}\{\frac{\partial^4 m(\mathbf{x})}{\partial x_i^3 \partial x_j} + \frac{\partial^4 m(\mathbf{x})}{\partial x_i \partial x_j^3}\} + \frac{1}{2}\mu_2\sum_{\substack{1\le k\le p \\ k\ne i,j}}\frac{\partial^4 m(\mathbf{x})}{\partial x_k^2 \partial x_i \partial x_j}, \\ & \textit{for } 1 \le j < i \le p. \end{aligned}$$

and

$$\gamma_1(m,K) = (\gamma_{11}(1), \cdots, \gamma_{p1}(1), \gamma_{22}(1), \cdots, \gamma_{p2}(1), \cdots, \gamma_{pp}(1))^T.$$

$$\begin{aligned}\gamma_{ii}(1) &= \frac{1}{3!}\frac{\mu_2\mu_6-\mu_4^2}{\mu_2(\mu_4-\mu_2^2)}\frac{\partial^3 m(\mathbf{x})}{\partial x_i^3}\frac{\partial f(\mathbf{x})}{\partial x_i}+\frac{1}{2}\mu_2\sum_{\substack{1\le k\le p\\ k\ne i}}\frac{\partial^3 m(\mathbf{x})}{\partial x_i^2\partial x_k}\frac{\partial f(\mathbf{x})}{\partial x_k},\\ & \textit{for } 1\le i\le p.\\ \gamma_{ij}(1) &= \mu_2\sum_{\substack{1\le k\le p\\ k\ne i,j}}\frac{\partial^3 m(\mathbf{x})}{\partial x_k\partial x_i\partial x_j}\frac{\partial f(\mathbf{x})}{\partial x_k}-\frac{\mu_2}{2}\{\frac{\partial^3 m(\mathbf{x})}{\partial x_i\partial x_j^2}\frac{\partial f(\mathbf{x})}{\partial x_j}+\frac{\partial^3 m(\mathbf{x})}{\partial x_i^2\partial x_j}\frac{\partial f(\mathbf{x})}{\partial x_i}\},\\ & \textit{for } 1\le j< i\le p.\end{aligned}$$

The conditional variance-covariance matrix is given by:

$$\begin{aligned}&\mathrm{Cov}\left\{\left(\begin{array}{c}\widehat{m(\mathbf{x})}\\ \widehat{D_m(\mathbf{x})}\\ \mathrm{vech}\{\widehat{L(\mathbf{x})}\}\end{array}\right)\middle| X_1,X_2,\ldots,X_n\right\}=\\ &\frac{\nu(\mathbf{x})}{nh^pf(\mathbf{x})}\left(\begin{array}{ccc}\rho & 0 & \phi h^{-2}\mathrm{vech}^T\{I\}\\ 0 & J_2\mu_2^{-2}h^{-2}I & 0\\ \phi h^{-2}\mathrm{vech}\{I\} & 0 & h^{-4}\{\Lambda-\frac{\mu_2(J_2-J_0\mu_2)}{(\mu_4-\mu_2^2)^2}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}\}\end{array}\right)\\ &+\frac{1}{nh^p}\mathrm{diag}\{1,h^{-1}I_1,h^{-2}I_2\}[o(1)+O_p((nh^p)^{-\frac{1}{2}})]\mathrm{diag}\{1,h^{-1}I_1,h^{-2}I_2\},\qquad (3.12)\end{aligned}$$

where

$$\rho=(\mu_4-\mu_2^2)^{-2}\{J_0(\mu_4+(p-1)\mu_2^2)^2-2pJ_2\mu_2(\mu_4+(p-1)\mu_2^2)+p\mu_2^2(J_4+(p-1)J_2^2)\},$$

$$\phi=(\mu_4-\mu_2^2)^{-2}\{J_2\mu_4+(2p-1)J_2\mu_2^2-(p-1)J_2^2\mu_2-J_4\mu_2-J_0\mu_2\mu_4-(p-1)J_0\mu_2^3\},$$

$$\Lambda=\mathrm{diag}\{\lambda_1,\underbrace{\lambda_2,\cdots,\lambda_2}_{p-1},\lambda_1,\underbrace{\lambda_2,\cdots,\lambda_2}_{p-2},\cdots,\lambda_1,\lambda_2,\lambda_1\},$$

where

$$\lambda_1=(J_4-J_2^2)(\mu_4-\mu_2^2)^{-2},\lambda_2=J_2^2\mu_2^{-4}.$$

Note that μ_ℓ, J_ℓ are defined in Theorem 3.1 as moments of K and K^2, respectively.

REMARKS ON THEOREM 3.2:

1. For the results on the first-order partial derivative part to hold, the assumptions $m\in C^4(U), f\in C^1(U)$ are not necessary. Instead, the weaker assumptions $m\in C^3(U)$ $f\in C^0(U)$ will suffice.

2. The assumptions $m \in C^4(U), f \in C^1(U)$ are necessary for the regression estimator and the estimators of second-order (mixed) partial derivatives, i.e. Hessian matrix estimation. Under weaker assumption $m \in C^3(U), f \in C^{(0)}(U)$, their bias have lower orders of $O(h^3)$ and $O(h)$, respectively.

3. We can compare the performance of first-order partial derivative estimator from the local quadratic fit to that from the local linear fit in Theorem 3.1 of Section 3.3 (under the same smoothness assumption on the regression function). It is seen that the local quadratic approach eliminates the undesirable bias term $\frac{h^2}{2\mu_2 f(\mathbf{x})} b_1(m, K)$ in the local linear derivative estimator.

4. It can be shown that the order of bias and covariance matrix for the first-order partial derivative estimator at the boundary remains the same, i.e. the locally quadratic polynomial approach to estimating first-order partial derivatives is free from the boundary effect in the sense that the asymptotic convergence rate is unaffected at the boundary, though more variability may result due to fewer observations being used. On the other hand it can be checked that the local linear estimator of partial derivatives suffers from boundary effects, in that the order of bias at the boundary points is of $O(h)$ instead of $O(h^2)$.

The following corollary discusses the properties of the local quadratic partial derivative estimator as a consequence of Theorem 3.2

Corollary 3.1 *The pointwise conditional mean squared error (CMSE) of the gradient vector $D_m(\mathbf{x})$ is given by*

$$E\{\|\widehat{D_m(\mathbf{x})} - D_m(\mathbf{x})\|^2 \mid X_1, X_2, \cdots, X_n\} \approx \frac{h^4}{(3!\mu_2)^2}\|b(m, K)\|^2 + \frac{pJ_2\nu(\mathbf{x})}{\mu_2^2 n h^{p+2} f(\mathbf{x})}. \tag{3.13}$$

The locally optimal h which minimizes (3.13) is given by

$$h_{\text{opt}}(\mathbf{x}) = \{\frac{9p(p+2)J_2\nu(\mathbf{x})}{f(\mathbf{x})\|b(m, K)\|^2}\}^{\frac{1}{p+6}} n^{-\frac{1}{p+6}}.$$

The minimum pointwise CMSE is given by plugging in $h_{\text{opt}}(\mathbf{x})$

$$\text{CMSE}_{\text{opt}}(\mathbf{x}) = \frac{\{p(p+2)J_2\nu(\mathbf{x})\}^{\frac{4}{p+6}}\|b(m,K)\|^{\frac{2p+4}{p+6}}}{3^{\frac{2p+4}{p+6}}4\mu_2^2 f(\mathbf{x})^{\frac{4}{p+6}}} n^{-\frac{4}{p+6}}.$$

It is seen from Corollary 3.1 that the convergence rate of the estimator with the optimal h is given by $n^{-\frac{2}{p+6}}$, which attains the optimal rate as established in Stone (1980). The asymptotic analysis provides some insights on the behavior of the estimator. The bias is quantified by the amount of smoothing and the third-order partial derivatives at x for each coordinate. Bias is increased when there is more third-order nonlinearity given by $b(m, K)$ and more smoothing. On the other hand, the conditional variance will be increased when there is less smoothing and sparser data.

3.5 Discussion

The local polynomial fit can also be applied to estimate other conditional functionals rather than the conditional mean. Examples include the conditional variance and the conditional distribution function. When the conditional distribution is asymmetric, percentile regression may be more informative, which also has the robustness property. The idea of local polynomial fit is also useful for density estimation.

Another research problem is data-based bandwidth selection of the regression estimator, for which the result on the Hessian matrix estimation is particularly useful. The local polynomial method can also be used in time series analysis, e.g. in model identification and nonlinear prediction. Results in this chapter will be generalized to time series in Chapter 4.

In principle, generalization to a higher-order polynomial fit can be considered similarly. However, there are serious limitations to higher-order local polynomial fits. The local polynomial method suffers from the usual "curse of dimensionality" problem in nonparametric regression estimators, that is, the data required for a given precision goes up exponentially as the dimension p of predictor variables increases. The problem gets worse for higher-order fit, such as third-order, or fourth-order polynomial fit.

In some situations, the global estimation of a regression or partial derivative is of interest. Since the regression or derivative surface may be smoother at some parts, and rougher at other parts, while the design density may be higher at some parts and lower at other parts, it is certainly advantageous to require an estimator to be spatially adaptive in the sense it chooses the bandwidth locally to adapt to local conditions. By using ideas of variable bandwidth and nearest neighbor estimation, the estimators considered here can be modified to have the adaptation property. Consult Fan and Gijbels (1992) for results in the univariate case.

Other approaches have been proposed in the literature to deal with high-dimensional data, including CART, Projection Pursuit, MARS, and neural net. CART and MARS also possess the adaptation property. Essentially, these methods search for a lower-dimensional representation of the underlying data, and hence avoiding the dimensionality problem. In principle, the local polynomial fit can be used in combination with CART, Projection Pursuit, and MARS to reduce the curse of dimensionality.

Though polynomial functions provide a natural local representation for smooth functions based on the principle of Taylor expansion, there are certainly some other representation functions which are worth considering in some situations, such as the wavelets basis. The wavelets estimator is nonlinear, in the sense that it is a nonlinear function of $\{Y_i\}$, as opposed to the linear estimator, which can be written as a linear combination of the Y's with the weights depending only on the X's, e.g. the local polynomial estimators are linear. In some situations, linear estimators cannot be improved, and nonlinear estimators such as wavelets may do better. Furthermore, the wavelets estimator automatically has the adaptation property which makes it attractive. However, since the local polynomial method is easy to interpret, to study, and to implement, and many theoretical properties are known about it, the local polynomial approach will remain a popular choice among so many nonparametric smoothing methods.

3.6 Proofs

Theorem 3.1 is considered easier to prove and follows along the lines of Theorem 3.2. So we only prove the latter here.

Some notations. We introduce some notations to simplify expressions in (3.9) and (3.10).

Rewrite

$$\mathbf{X} = \mathrm{diag}\{1, hI_1, h^2 I_2\} \begin{pmatrix} 1 & (\frac{X_1-\mathbf{x}}{h})^T & \mathrm{vech}^T\{(\frac{X_1-\mathbf{x}}{h})(\frac{X_1-\mathbf{x}}{h})^T\} \\ \vdots & \vdots & \vdots \\ 1 & (\frac{X_n-\mathbf{x}}{h})^T & \mathrm{vech}^T\{(\frac{X_n-\mathbf{x}}{h})(\frac{X_n-\mathbf{x}}{h})^T\} \end{pmatrix},$$

where I_1 and I_2 are identity matrices of dimension p and $p(p+1)/2$, respectively. Then,

$$\frac{1}{nh^p}\mathbf{X}^T W \mathbf{X} = \mathrm{diag}\{1, hI_1, h^2 I_2\} \cdot S_n \cdot \mathrm{diag}\{1, hI_1, h^2 I_2\},$$

where we denote

$$S_n = \frac{1}{nh^p}\sum_{i=1}^{n} S(n,i), \tag{3.14}$$

where

$$S(n,i) = \begin{pmatrix} s_{11}(n,i) & s_{21}^T(n,i) & s_{31}^T(n,i) \\ s_{21}(n,i) & s_{22}(n,i) & s_{32}^T(n,i) \\ s_{31}(n,i) & s_{32}(n,i) & s_{33}(n,i) \end{pmatrix}$$

whose components are given by

$$s_{11}(n,i) = K(\frac{X_i-\mathbf{x}}{h}), s_{21}(n,i) = (\frac{X_i-\mathbf{x}}{h})K(\frac{X_i-\mathbf{x}}{h}),$$

$$s_{31}(n,i) = \mathrm{vech}\{(\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\}K(\frac{X_i-\mathbf{x}}{h}),$$

$$s_{22}(n,i) = (\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T K(\frac{X_i-\mathbf{x}}{h}),$$

$$s_{32}(n,i) = \mathrm{vech}\{(\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\}(\frac{X_i-\mathbf{x}}{h})^T K(\frac{X_i-\mathbf{x}}{h}),$$

$$s_{33}(n,i) = \mathrm{vech}\{(\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\}\mathrm{vech}^T\{(\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\}K(\frac{X_i-\mathbf{x}}{h}).$$

Also rewrite

$$\frac{1}{nh^p}\mathbf{X}^T W(M-\mathbf{X}\beta) = \mathrm{diag}\{1, hI_1, h^2 I_2\}R_n,$$

where we denote

$$R_n = \frac{1}{nh^p}\sum_{i=1}^{n} R(n,i), \text{ and } R(n,i) = (r_1(n,i), r_2(n,i), r_3(n,i))^T, \tag{3.15}$$

with its components given by

$$r_1(n,i) = \{m(X_i) - m(\mathbf{x}) - D_m^T(\mathbf{x})(X_i - \mathbf{x}) - \frac{1}{2}(X_i-\mathbf{x})^T H_m(\mathbf{x})(X_i-\mathbf{x})\}K(\frac{X_i-\mathbf{x}}{h}),$$

$$\begin{aligned} r_2(n,i) &= (\frac{X_i-\mathbf{x}}{h})\{m(X_i) - m(\mathbf{x}) - D_m^T(\mathbf{x})(X_i-\mathbf{x}) \\ &\quad -\frac{1}{2}(X_i-\mathbf{x})^T H_m(\mathbf{x})(X_i-\mathbf{x})\}K(\frac{X_i-\mathbf{x}}{h}), \\ r_3(n,i) &= \mathrm{vech}\{(\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\}\{m(X_i) - m(\mathbf{x}) - D_m^T(\mathbf{x})(X_i-\mathbf{x}) \\ &\quad -\frac{1}{2}(X_i-\mathbf{x})^T H_m(\mathbf{x})(X_i-\mathbf{x})\}K(\frac{X_i-\mathbf{x}}{h}). \end{aligned}$$

Similarly,

$$\frac{1}{nh^p}\mathbf{X}^T WVW\mathbf{X} = \mathrm{diag}\{1, hI_1, h^2 I_2\}C_n \mathrm{diag}\{1, hI_1, h^2 I_2\},$$

where denote

$$C_n = \frac{1}{nh^p}\sum_{i=1}^{n} C(n,i), \tag{3.16}$$

$$C(n,i) = \begin{pmatrix} c_{11}(n,i) & c_{21}^T(n,i) & c_{31}^T(n,i) \\ c_{21}(n,i) & c_{22}(n,i) & c_{32}^T(n,i) \\ c_{31}(n,i) & c_{32}(n,i) & c_{33}(n,i) \end{pmatrix}$$

whose components are given by

$$c_{11}(n,i) = v(X_i)K^2(\frac{X_i-\mathbf{x}}{h}), c_{21}(n,i) = (\frac{X_i-\mathbf{x}}{h})\nu(X_i)K^2(\frac{X_i-\mathbf{x}}{h}),$$

$$c_{31}(n,i) = \mathrm{vech}\{(\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\}\nu(X_i)K^2(\frac{X_i-\mathbf{x}}{h}),$$

$$c_{22}(n,i) = (\frac{X_i-\mathbf{x}}{h})(\frac{X_i-\mathbf{x}}{h})^T\nu(X_i)K^2(\frac{X_i-\mathbf{x}}{h}),$$

$$c_{32}(n,i) = \text{vech}\{(\frac{X_i - \mathbf{x}}{h})(\frac{X_i - \mathbf{x}}{h})^T\}(\frac{X_i - \mathbf{x}}{h})^T \nu(X_i) K^2(\frac{X_i - \mathbf{x}}{h}),$$

$$c_{33}(n,i) = \text{vech}\{(\frac{X_i - \mathbf{x}}{h})(\frac{X_i - \mathbf{x}}{h})^T\}\text{vech}^T\{(\frac{X_i - \mathbf{x}}{h})(\frac{X_i - \mathbf{x}}{h})^T\} \nu(X_i) K^2(\frac{X_i - \mathbf{x}}{h}).$$

Note that the argument of n in $S(n,i), R(n,i), C(n,i)$ denotes the dependence on sample size n through h, which tends to zero proportional to n. Using the introduced notations, the conditional bias of $\hat{\beta}$ can be re-expressed as:

$$E(\hat{\beta} \mid X_1, \cdots, X_n) - \beta = \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\} S_n^{-1} R_n. \tag{3.17}$$

The conditional variance-covariance matrix of the estimator can be rewritten as:

$$\begin{aligned} &\text{Cov}(\hat{\beta} \mid X_1, \cdots, X_n) = \\ &\quad \frac{1}{nh^p} \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\} S_n^{-1} C_n S_n^{-1} \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}. \end{aligned} \tag{3.18}$$

Some preliminaries on stochastic order. Given matrices $A_n = A + O_p(\alpha_n), B_n = B + O_p(\gamma_n), \alpha_n \to 0, \gamma_n \to 0$ we have

$$A_n B_n = AB + O_p(\max\{\alpha_n, \gamma_n\}). \tag{3.19}$$

We also have the following proposition on the inverse of a random matrix.

Proposition 3.1 *Let $\{A_n\}$ be a sequence of random $p \times p$ matrices such that*

$$A_n - A = O_p(\gamma_n),$$

where A is a constant $p \times p$ matrix and $\gamma_n \to 0$ as $n \to 0$. If A is invertible, then

$$A_n^{-1} = A^{-1} + O_p(\gamma_n). \tag{3.20}$$

Proof: Note that by the matrix differential $dA^{-1} = -A^{-1} \cdot dA \cdot A^{-1}$, we have

$$A_n^{-1} - A^{-1} = A^{-1}(A_n - A)A^{-1} + o(\|A_n - A\|).$$

It follows that,

$$A_n^{-1} - A^{-1} = O_p(\gamma_n)$$

by assumption. □

If a random sequence $\{Z_n\}$ has second moments, $EZ_n^2 < \infty$ for every n, a natural stochastic order is given by its standard deviation, that is

$$Z_n = EZ_n + O_p(\{\mathrm{Var}(Z_n)\}^{\frac{1}{2}}). \tag{3.21}$$

This can be shown to follow easily from the Chebyshev's inequality.

Since sums of triangular arrays are frequently encountered later on, we have the following proposition as variations of (3.21).

Proposition 3.2 *Given random variables $Z_{n1}, \dots, Z_{nn}$ which are defined on the same probability space for each $n = 1, 2, \dots$, and which satisfy*

$$EZ_{nk}^2 < \infty, \ \textit{for } 1 \leq k \leq n, n = 1, 2, \dots$$

we have

$$\sum_{k=1}^{n} Z_{nk} = \sum_{k=1}^{n} EZ_{nk} + O_p(\{\mathrm{Var}(\sum_{k=1}^{n} Z_{nk})\}^{\frac{1}{2}}). \tag{3.22}$$

In particular, if $Z_{n1}, \dots, Z_{nn}$ are iid for each n,

$$\sum_{k=1}^{n} Z_{nk} = nEZ_{n1} + O_p(\{n\mathrm{Var}Z_{n1}\}^{\frac{1}{2}}). \tag{3.23}$$

We now give several lemmas which are needed in our proofs. The behaviors of S_n^{-1} and R_n which appear in the conditional bias of (3.17) are studied in Lemma 3.1-3.5.

Several lemmas.

Lemma 3.1 *For S_n given in (3.14), as $nh^p \to \infty$, we have*

$$S_n = A(h) + O_p(\{nh^p\}^{-\frac{1}{2}}). \tag{3.24}$$

where

$$A(h) = \begin{pmatrix} a_{11}(h) & a_{21}^T(h) & a_{31}^T(h) \\ a_{21}(h) & a_{22}(h) & a_{32}^T(h) \\ a_{31}(h) & a_{32}(h) & a_{33}(h) \end{pmatrix}, \tag{3.25}$$

and

$$\begin{aligned}
a_{11}(h) &= \int K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u},\\
a_{21}(h) &= \int \mathbf{u}K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u},\\
a_{31}(h) &= \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u},\\
a_{22}(h) &= \int \mathbf{u}\mathbf{u}^T K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u},\\
a_{32}(h) &= \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}\mathbf{u}^T K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u},\\
a_{33}(h) &= \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}\mathrm{vech}^T\{\mathbf{u}\mathbf{u}^T\}K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u}.
\end{aligned}$$

Proof: Recall that $S_n = (nh^p)^{-1}\sum_{i=1}^n S(n,i)$ and $S(n,i)$'s are iid random matrix for each n. Note that

$$ES(n,1) = h^p A(h),$$

by change of variables $\mathbf{u} = (\mathbf{x}_1 - \mathbf{x})/h$ in the integrals. Similarly, it can be shown that the variance of each element in $S(n,1)$ has order $O(h^p)$. Apply equation (3.23) in Proposition 3.2 to each element of S_n, we have

$$S_n = h^{-p}ES(n,1) + O_p(\{nh^p\}^{\frac{1}{2}}) = A(h) + O_p(\{nh^p\}^{-\frac{1}{2}}).$$

The lemma is verified. □

By letting $h \to 0$, $A(h)$ in (3.25) can be simplified. By substituting Taylor expansion for $f \in C^1(U)$

$$f(\mathbf{x}+h\mathbf{u}) = f(\mathbf{x}) + hD_f^T(\mathbf{x})\mathbf{u} + o(h), \text{as } h \to 0,$$

we obtain the following lemma.

Lemma 3.2 *For an interior point* $\mathbf{x}$*, if there exists a neighborhood* U *satisfying* $f \in$

$C^1(U)$, then as $h \to 0$

$$
\begin{aligned}
A(h) &= f(\mathbf{x}) \begin{pmatrix} 1 & 0 & \mu_2 \mathrm{vech}^T\{I\} \\ 0 & \mu_2 I & 0 \\ \mu_2 \mathrm{vech}\{I\} & 0 & D \end{pmatrix} \\
& \quad + h \begin{pmatrix} 0 & \mu_2 D_f^T(\mathbf{x}) & 0 \\ \mu_2 D_f(\mathbf{x}) & 0 & H \\ 0 & H^T & 0 \end{pmatrix} + o(h),
\end{aligned}
$$

where

$$
\begin{aligned}
D &= \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\} \mathrm{vech}^T\{\mathbf{u}\mathbf{u}^T\} K(\mathbf{u}) d\mathbf{u}, \\
&= E + \mu_2^2 \mathrm{vech}\{I\} \mathrm{vech}^T\{I\}, \qquad (3.26)
\end{aligned}
$$

where we define

$$
E = \mathrm{diag}\{\mu_4 - \mu_2^2, \underbrace{\mu_2^2, \cdots, \mu_2^2}_{p-1}, \mu_4 - \mu_2^2, \underbrace{\mu_2^2, \cdots, \mu_2^2}_{p-2}, \cdots, \mu_4 - \mu_2^2, \mu_2^2, \mu_4 - \mu_2^2\}. \quad (3.27)
$$

and

$$
H = \int \mathbf{u} \mathrm{vech}^T\{\mathbf{u}\mathbf{u}^T\} K(\mathbf{u}) D_f^T(\mathbf{x}) \mathbf{u} d\mathbf{u} =
$$

$$
\mu_2^2 \begin{pmatrix}
\frac{\partial f(\mathbf{x})}{\partial x_1} \frac{\mu_4}{\mu_2^2} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_p} & \frac{\partial f(\mathbf{x})}{\partial x_1} & 0 & \cdots & 0 & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_1} & 0 & \frac{\partial f(\mathbf{x})}{\partial x_1} \\
\frac{\partial f(\mathbf{x})}{\partial x_2} & \frac{\partial f(\mathbf{x})}{\partial x_1} & \cdots & 0 & \frac{\partial f(\mathbf{x})}{\partial x_2} \frac{\mu_4}{\mu_2^2} & \frac{\partial f(\mathbf{x})}{\partial x_3} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_p} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_2} & 0 & \frac{\partial f(\mathbf{x})}{\partial x_2} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots \\
\frac{\partial f(\mathbf{x})}{\partial x_p} & 0 & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_p} & 0 & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_p} & \frac{\partial f(\mathbf{x})}{\partial x_{p-1}} & \frac{\partial f(\mathbf{x})}{\partial x_p} \frac{\mu_4}{\mu_2^2}
\end{pmatrix}.
$$

Proof: It can be verified that D, E, and H have the given forms in the lemma. □

Now let's make some calculations for matrices in Lemma 3.2. By using a formula for matrix inverse (e.g. see problem 2.8 of p33 Rao (1973)), and the relation

$$
E^{-1} \mathrm{vech}\{I\} = (\mu_4 - \mu_2^2)^{-1} \mathrm{vech}\{I\}, \qquad (3.28)
$$

the inverse of matrix D in (3.26) is given by

$$
D^{-1} = E^{-1} - \frac{\mu_2^2}{(\mu_4 - \mu_2^2)(\mu_4 + (p-1)\mu_2^2)} \mathrm{vech}\{I\} \mathrm{vech}^T\{I\},
$$

where

$$\begin{aligned} E^{-1} \;=\;& \mathrm{diag}\{(\mu_4-\mu_2^2)^{-1}, \underbrace{\mu_2^{-2},\cdots,\mu_2^{-2}}_{p-1}, (\mu_4-\mu_2^2)^{-1}, \underbrace{\mu_2^{-2},\cdots,\mu_2^{-2}}_{p-2}, \\ & \cdots, (\mu_4-\mu_2^2)^{-1}, \mu_2^{-2}, (\mu_4-\mu_2^2)^{-1}\}. \end{aligned}$$

Next, we use

$$\begin{pmatrix} 1 & b^T \\ b & C \end{pmatrix}^{-1} = \frac{1}{d}\begin{pmatrix} 1 & -b^TC^{-1} \\ -C^{-1}b & dC^{-1}+C^{-1}bb^TC^{-1} \end{pmatrix},$$

where b is a vector, while C is a symmetric matrix and C^{-1} exists, and $d = 1-b^TC^{-1}b$.

Applying it to the leading matrix term of $A(h)$ given by

$$\begin{pmatrix} 1 & 0 & \mu_2\mathrm{vech}^T\{I\} \\ 0 & \mu_2 I & 0 \\ \mu_2\mathrm{vech}\{I\} & 0 & D \end{pmatrix},$$

by identifying

$$b = \begin{pmatrix} 0 \\ \mu_2\mathrm{vech}\{I\} \end{pmatrix}, C = \begin{pmatrix} \mu_2 I & 0 \\ 0 & D \end{pmatrix}. \tag{3.29}$$

Note that

$$C^{-1} = \begin{pmatrix} \mu_2^{-1}I & 0 \\ 0 & E^{-1} - \frac{\mu_2^2}{(\mu_4-\mu_2^2)(\mu_4+(p-1)\mu_2^2)}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\} \end{pmatrix}, \tag{3.30}$$

we have

$$\begin{aligned} C^{-1}b &= \begin{pmatrix} 0 \\ \mu_2E^{-1}\mathrm{vech}\{I\} - \frac{p\mu_2^3}{(\mu_4-\mu_2^2)(\mu_4+(p-1)\mu_2^2)}\mathrm{vech}\{I\} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \frac{\mu_2}{\mu_4+(p-1)\mu_2^2}\mathrm{vech}\{I\} \end{pmatrix}, \end{aligned}$$

by using the relation (3.28).

Further,

$$\begin{aligned}
d &= 1 - b^T C^{-1} b \\
&= 1 - \mu_2^2 \mathrm{vech}^T\{I\} D^{-1} \mathrm{vech}\{I\} \\
&= 1 - \mu_2^2 \mathrm{vech}^T\{I\} E^{-1} \mathrm{vech}\{I\} + \frac{\mu_2^4}{(\mu_4 - \mu_2^2)(\mu_4 + (p-1)\mu_2^2)} (\mathrm{vech}^T\{I\}\mathrm{vech}\{I\})^2 \\
&= 1 - \frac{p\mu_2^2}{\mu_4 - \mu_2^2} + \frac{p^2\mu_2^4}{(\mu_4 - \mu_2^2)(\mu_4 + (p-1)\mu_2^2)},
\end{aligned}$$

after simplification,

$$d = (\mu_4 - \mu_2^2)/(\mu_4 + (p-1)\mu_2^2). \tag{3.31}$$

Note also that

$$C^{-1}bb^TC^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & \mu_2^2 D^{-1}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}D^{-1} \end{pmatrix}.$$

Using (3.28), it can be shown that

$$D^{-1}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\} = (\mu_4 + (p-1)\mu_2^2)^{-1}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}.$$

Noticing $\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}D^{-1} = D^{-1}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}$, we have

$$C^{-1}bb^TC^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{\mu_2^2}{(\mu_4 + (p-1)\mu_2^2)^2}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\} \end{pmatrix}.$$

Combined with formulas (3.30) of C^{-1} and (3.31) of d, we obtain

$$dc^{-1} + C^{-1}bb^TC^{-1} = \begin{pmatrix} d\mu_2^{-1}I & 0 \\ 0 & dE^{-1} \end{pmatrix}.$$

In all, the inverse of (3.6) is given by

$$\begin{pmatrix} d^{-1} & 0 & -\mu_2(\mu_4 - \mu_2^2)^{-1}\mathrm{vech}^T\{I\} \\ 0 & \mu_2^{-1}I & 0 \\ -\mu_2(\mu_4 - \mu_2^2)^{-1}\mathrm{vech}\{I\} & 0 & E^{-1} \end{pmatrix},$$

where d is given in (3.31).

Now we can give the Taylor expansion of $A(h)^{-1}$ in next lemma.

Lemma 3.3 *Under assumptions of lemma 3.2 and $f(\mathbf{x}) > 0$, we have as $h \to 0$:*

$$
\begin{aligned}
A^{-1}(h) \;=\; & \frac{1}{f(\mathbf{x})}\begin{pmatrix} d^{-1} & 0 & -\mu_2(\mu_4-\mu_2^2)^{-1}\mathrm{vech}^T\{I\} \\ 0 & \mu_2^{-1}I & 0 \\ -\mu_2(\mu_4-\mu_2^2)^{-1}\mathrm{vech}\{I\} & 0 & E^{-1} \end{pmatrix} \\
& -\frac{h}{\mu_2 f(\mathbf{x})^2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & N \\ 0 & N^T & 0 \end{pmatrix} + o(h), \qquad (3.32)
\end{aligned}
$$

where $d = (\mu_4 - \mu_2^2)/(\mu_4 + (p-1)\mu_2^2)$, N is a $p \times (p(p+1)/2)$ matrix, and is defined as

$$
\begin{pmatrix}
\frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_{p-1}} & \frac{\partial f(\mathbf{x})}{\partial x_p} & & & & & & & & \\
& \frac{\partial f(\mathbf{x})}{\partial x_1} & & & & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_{p-1}} & \frac{\partial f(\mathbf{x})}{\partial x_p} & & & & \\
& & \ddots & & & & \ddots & & & \cdots & & & \\
& & & \frac{\partial f(\mathbf{x})}{\partial x_1} & & & & \frac{\partial f(\mathbf{x})}{\partial x_2} & & & \frac{\partial f(\mathbf{x})}{\partial x_{p-1}} & \frac{\partial f(\mathbf{x})}{\partial x_p} & \\
& & & & \frac{\partial f(\mathbf{x})}{\partial x_1} & & & & \frac{\partial f(\mathbf{x})}{\partial x_2} & & & \frac{\partial f(\mathbf{x})}{\partial x_{p-1}} & \frac{\partial f(\mathbf{x})}{\partial x_p}
\end{pmatrix}
$$

where the suppressed elements are zeroes.

Proof: The first-order term $A(0)^{-1}$ has already been shown to have the given form. By employing the matrix differential

$$dA^{-1} = -A^{-1} \cdot dA \cdot A^{-1},$$

the second term in (3.32) is given by

$$
\begin{aligned}
& -\frac{h}{f^2(\mathbf{x})}\begin{pmatrix} d^{-1} & 0 & -cV^T \\ 0 & \mu_2^{-1}I & 0 \\ -cV & 0 & E^{-1} \end{pmatrix}\begin{pmatrix} 0 & \mu_2 D_f^T & 0 \\ \mu_2 2D_f & 0 & H \\ 0 & H^T & 0 \end{pmatrix}\begin{pmatrix} d^{-1} & 0 & -cV^T \\ 0 & \mu_2^{-1}I & 0 \\ -cV & 0 & E^{-1} \end{pmatrix} \\
& = -\frac{h}{f^2(\mathbf{x})}\begin{pmatrix} 0 & d^{-1}D_f^T - c\mu_2^{-1}V^T H^T & 0 \\ d^{-1}D_f - c\mu_2^{-1}HV & 0 & -cD_f V^T + \mu_2^{-1}HE^{-1} \\ 0 & -cVD_f^T + \mu_2^{-1}E^{-1}H^T & 0 \end{pmatrix},
\end{aligned}
$$

here $V = \text{vech}\{I\}$ and $c = \frac{\mu_2}{\mu_4 - \mu_2^2}$.

Note that

$$HV = (\mu_4 + (p-1)\mu_2^2)D_f,$$

thus

$$\frac{c}{\mu_2}HV = \frac{\mu_4 + (p-1)\mu_2^2}{\mu_4 - \mu_2^2}D_f = \frac{1}{d}D_f,$$

i.e.

$$d^{-1}D_f - c\mu_2^{-1}HV = 0,$$

$$d^{-1}D_f^T - c\mu_2^{-1}V^T H^T = 0.$$

On the other hand, by using the explicit expression for H given in Lemma 3.2, it can be easily checked that

$$N = HE^{-1} - \mu_2 c D_f V^T$$

has the form given in the lemma. □

In summary, we have the following lemma on S_n^{-1}.

Lemma 3.4 *For an interior point* $\mathbf{x}$ *with* $f(\mathbf{x}) > 0$*, if there exists an open neighborhood* U *such that* $f \in C^1(U)$*, we have that as* $h \to 0, nh^p \to \infty$*,*

$$S_n^{-1} = A^{-1}(h) + O_p(\{nh^p\}^{-\frac{1}{2}}), \tag{3.33}$$

where $A(h)^{-1}$ *has expansions as given in Lemma 3.3.*

Proof: Using Lemma 3.1, Lemma 3.3, and applying Proposition 3.1 to S_n, we have that, as $h \to 0, nh^p \to \infty$,

$$S_n^{-1} = A^{-1}(h) + O_p((nh^p)^{-\frac{1}{2}}).$$

□

Next, we consider the behavior of R_n which is defined in (3.15). We have the following lemma.

Lemma 3.5 *Assume that K is spherically symmetric and satisfies*

$$\int u_1^{12} K(u_1, \cdots, u_p) du_1 \cdots du_p < \infty,$$

and that $m \in C^4(U), f \in C^1(U)$, then as $h \to 0, nh^p \to \infty$,

$$R_n = h^3\{R(h, \mathbf{x}) + o(h) + O_p(\{nh^p\}^{-\frac{1}{2}})\}, \tag{3.34}$$

where $R(h, \mathbf{x})$ is defined by

$$R(h, \mathbf{x}) = \frac{1}{3!} \begin{pmatrix} h \int (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(\mathbf{u}) D_f^T(\mathbf{x}) \mathbf{u} d\mathbf{u} \\ f(\mathbf{x}) \int \mathbf{u} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(\mathbf{u}) d\mathbf{u} \\ h \int \text{vech}\{\mathbf{u}\mathbf{u}^T\} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(\mathbf{u}) D_f^T(\mathbf{x}) \mathbf{u} d\mathbf{u} \end{pmatrix}$$

$$+ \frac{f(\mathbf{x})h}{4!} \begin{pmatrix} \int (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^4 m(\mathbf{x}) K(\mathbf{u}) d\mathbf{u} \\ 0 \\ \int \text{vech}\{\mathbf{u}\mathbf{u}^T\} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^4 m(\mathbf{x}) K(\mathbf{u}) d\mathbf{u} \end{pmatrix}. \tag{3.35}$$

Proof: Note that

$$\begin{aligned} ER(n,1) &= (Er_1(n,1), Er_2(n,1), Er_3(n,1))^T \\ &= h^p \int_{\mathcal{R}^p} (1, \mathbf{u}, \text{vech}\{\mathbf{u}\mathbf{u}^T\})^T \\ &\quad \{m(\mathbf{x} + h\mathbf{u}) - m(\mathbf{x}) - hD_m^T(\mathbf{x})\mathbf{u} - \frac{1}{2}h^2 \mathbf{u}^T H_m(\mathbf{x}) \mathbf{u}\} K(\mathbf{u}) f(\mathbf{x} + h\mathbf{u}) d\mathbf{u}, \end{aligned}$$

by using change of variable $\mathbf{u} = (\mathbf{x}_1 - \mathbf{x})/h$ in the integral.

Since $m \in C^{(4)}(U)$, we have the Taylor expansion for $m(\mathbf{x} + h\mathbf{u})$, as $h \to 0$

$$\begin{aligned} m(\mathbf{x} + h\mathbf{u}) &= m(\mathbf{x}) + hD_m(\mathbf{x})\mathbf{u} + \frac{1}{2}h^2 \mathbf{u}^T H_m \mathbf{u} + \frac{h^3}{3!} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) \\ &\quad + \frac{h^4}{4!} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^4 m(\mathbf{x}) + o(h^4). \end{aligned}$$

Substituting Taylor expansions for m and f, we have that

$$ER(n,1) = h^{p+3} \left[\frac{1}{3!} \begin{pmatrix} \int (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(\mathbf{u}) f(\mathbf{x} + h\mathbf{u}) d\mathbf{u} \\ \int \mathbf{u} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(\mathbf{u}) f(\mathbf{x} + h\mathbf{u}) d\mathbf{u} \\ \int \text{vech}\{\mathbf{u}\mathbf{u}^T\} (\sum_{i=1}^p \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(\mathbf{u}) f(\mathbf{x} + h\mathbf{u}) d\mathbf{u} \end{pmatrix} \right.$$

$$
\begin{aligned}
& +\frac{h}{4!}\begin{pmatrix} \int(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u} \\ \int \mathbf{u}\int(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u} \\ \int \text{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u} \end{pmatrix} + o(h)\Bigg] \\
= \; & h^{p+3}\Bigg[\frac{1}{3!}\begin{pmatrix} h\int(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})D_f^T(\mathbf{x})\mathbf{u}d\mathbf{u} \\ f(\mathbf{x})\int \mathbf{u}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \\ h\int \text{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})D_f^T(\mathbf{x})\mathbf{u}d\mathbf{u} \end{pmatrix} \\
& +\frac{f(\mathbf{x})h}{4!}\begin{pmatrix} \int(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \\ 0 \\ \int \text{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \end{pmatrix} + o(h)\Bigg], \\
= \; & h^{p+3}(R(h,\mathbf{x})+o(h)),
\end{aligned}
$$

where $R(h,\mathbf{x})$ is given in (3.35).

Similarly, it can be shown that

$$\text{Cov}\{R(n,1),R(n,1)\} = O(h^{p+6}).$$

By applying (3.23) in Proposition 3.2 to n iid q-dimensional random vectors $R(n,i)$'s, we obtain

$$R_n = h^3\{R(h,\mathbf{x})+o(h)+O_p((nh^p)^{-\frac{1}{2}})\}.$$

Lemma 3.5 is thus proved. □

Proof of bias part of Theorem 3.2.

Combining Lemma 3.5 with Lemma 3.3 and Lemma 3.4, we obtain the asymptotic expansion for the conditional bias of the estimator $\widehat{\beta}$ given by:

$$
\begin{aligned}
& E\{\widehat{\beta}-\beta \mid X_1,X_2,\cdots,X_n\} \\
= \; & \text{diag}\{1,h^{-1}I_1,h^{-2}I_2\}S_n^{-1}R_n \\
= \; & \text{diag}\{1,h^{-1}I_1,h^{-2}I_2\}[A^{-1}(h)+O_p(\{nh^p\}^{-\frac{1}{2}})] \\
& [h^3\{R(h,\mathbf{x})+o(h)+O_p(\{nh^p\}^{-\frac{1}{2}})\}] \\
= \; & h^3\text{diag}\{1,h^{-1}I_1,h^{-2}I_2\}\{A^{-1}(h)R(h,\mathbf{x})+o(h)+O_p(\{nh^p\}^{-\frac{1}{2}})\} \\
= \; & h^3\text{diag}\{1,h^{-1}I_1,h^{-2}I_2\}f(\mathbf{x})^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\left\{\begin{pmatrix} d^{-1} & 0 & -\frac{\mu_2}{\mu_4-\mu_2^2}\mathrm{vech}^T\{I\} \\ 0 & \mu_2^{-1}I & -\frac{h}{\mu_2 f(\mathbf{x})}N \\ -\frac{\mu_2}{\mu_4-\mu_2^2}\mathrm{vech}\{I\} & -\frac{h}{\mu_2 f(\mathbf{x})}N^T & E^{-1} \end{pmatrix}\right. \\
&\left[\frac{1}{3!}\begin{pmatrix} h\int(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})D_f^T(\mathbf{x})\mathbf{u}d\mathbf{u} \\ f(\mathbf{x})\int \mathbf{u}(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \\ h\int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})D_f^T(\mathbf{x})\mathbf{u}d\mathbf{u} \end{pmatrix}\right. \\
&\left.+\frac{f(\mathbf{x})h}{4!}\begin{pmatrix} \int(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \\ 0 \\ \int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \end{pmatrix}\right] \\
&\left.+o(h)+O_p(\{nh^p\}^{-\frac{1}{2}})\right\} \\
= &\begin{pmatrix} \frac{h^4}{4!}\theta(m,K)+\frac{h^4}{3!f(\mathbf{x})}\theta_1(m,K) \\ \frac{h^2}{3!\mu_2}b(m,K) \\ h^2\gamma(m,K)+\frac{h^2}{f(\mathbf{x})}\gamma_1(m,K) \end{pmatrix} \\
&+\mathrm{diag}\{h^3, h^2I_1, hI_2\}[o(h)+O_p(\{nh^p\}^{-\frac{1}{2}})],
\end{aligned}
$$

if $h \to 0, nh^p \to \infty$.

Here,

$$
b(m,K) = \int_{\mathcal{R}^p} u(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})d\mathbf{u},
$$

$$
\begin{aligned}
\theta(m,K) &= \frac{\mu_4+(p-1)\mu_2^2}{\mu_4-\mu_2^2}\int(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \\
&\quad -\frac{\mu_2}{\mu_4-\mu_2^2}\int(\sum_{i=1}^p u_i^2)(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u},
\end{aligned}
$$

$$
\begin{aligned}
\theta_1(m,K) &= \frac{\mu_4+(p-1)\mu_2^2}{\mu_4-\mu_2^2}\int(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})(D_f^T(\mathbf{x})\mathbf{u})d\mathbf{u} \\
&\quad -\frac{\mu_2}{\mu_4-\mu_2^2}\int(\sum_{i=1}^p u_i^2)(\sum_{i=1}^p \frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})(D_f^T(\mathbf{x})\mathbf{u})d\mathbf{u},
\end{aligned}
$$

by plugging in the value of d in Lemma 3.3.

$$
\begin{aligned}
\gamma(m,K) &= \frac{1}{4!}\{E^{-1}\int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u} \\
&\quad -\frac{\mu_2}{\mu_4-\mu_2^2}\mathrm{vech}\{I\}\int(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^4 m(\mathbf{x})K(\mathbf{u})d\mathbf{u}\},
\end{aligned}
$$

and

$$
\begin{aligned}
\gamma_1(m,K) &= \frac{1}{3!}\{E^{-1}\int \mathrm{vech}\{\mathbf{u}\mathbf{u}^T\}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})(D_f^T(\mathbf{x})\mathbf{u})d\mathbf{u} \\
&\quad -\frac{\mu_2}{\mu_4-\mu_2^2}\mathrm{vech}I\int(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})(D_f^T(\mathbf{x})\mathbf{u})d\mathbf{u} \\
&\quad -\mu_2^{-1}N^T\int \mathbf{u}(\sum_{i=1}^{p}\frac{\partial}{\partial x_i}u_i)^3 m(\mathbf{x})K(\mathbf{u})d\mathbf{u}\}.
\end{aligned}
$$

It can be checked that $\theta(m,K),\theta_1(m,K)$, and $\gamma(m,K),\gamma_1(m,K)$ have the forms given in the Theorem 3.2.

This completes the proof of the "bias" part of Theorem 3.2. □

Proof of covariance part of Theorem 3.2. The conditional covariance marix of $\hat{\beta}$ is given in (3.18). We need the following lemma on C_n defined in (3.16).

Lemma 3.6 *For an interior point* $\mathbf{x}$ *in the support of* f *such that* $f(\mathbf{x})>0,\nu(\mathbf{x})>0$. *If there exists an open neighborhood* U *such that* $\nu\in C^0(U), f\in C^1(U)$, *as* $h\to 0, nh^p\to\infty$, *we have*

$$C_n = C(\mathbf{x}) + O(h) + O_p((nh^p)^{-\frac{1}{2}}),$$

where

$$C(\mathbf{x}) = \nu(\mathbf{x})f(\mathbf{x})\begin{pmatrix} J_0 & 0 & J_2\mathrm{vech}^T\{I\} \\ 0 & J_2 I_1 & 0 \\ J_2\mathrm{vech}\{I\} & 0 & E_J + J_2^2\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}\end{pmatrix}, \qquad (3.36)$$

and

$$E_J = \mathrm{diag}\{J_4-J_2^2,\underbrace{J_2^2,\cdots,J_2^2}_{p-1},J_4-J_2^2,\underbrace{J_2^2,\cdots,J_2^2}_{p-2},\cdots,J_4-J_2^2,J_2^2,J_4-J_2^2\}. \quad (3.37)$$

Here $J_l=\int u_1^l K(\mathbf{u})^2 d\mathbf{u}$.

Proof: Going through similar calculations to S_n, and using Proposition 3.2, we have that

$$C_n = h^{-p}EC(n,1) + O_p((nh^p)^{-\frac{1}{2}}),$$

where

$$\begin{aligned} & EC(n,1) \\ = \ & h^p \nu(\mathbf{x}) f(\mathbf{x}) \begin{pmatrix} J_0 & 0 & J_2 \text{vech}^T\{I\} \\ 0 & J_2 I_1 & 0 \\ J_2 \text{vech}\{I\} & 0 & E_J + J_2^2 \text{vech}\{I\} vech^T\{I\} \end{pmatrix} + o(1) \\ = \ & h^p\{C(\mathbf{x}) + o(1)\}, \end{aligned}$$

where $C(\mathbf{x})$ is given in (3.36) of the lemma. The lemma is thus verified. □

Combining Lemma 3.6 with Lemma 3.4 and Lemma 3.3, we obtain the asymptotic expansion of the conditional covariance matrix as $h \to 0, nh^p \to \infty$.

$$\begin{aligned} & \text{Cov}(\hat{\beta} \mid X_1, X_2, \cdots, X_n) \\ = \ & \frac{1}{nh^p} \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}[A(h)^{-1}(C(\mathbf{x}) + o(1))A(h)^{-1} + O_p((nh^p)^{-\frac{1}{2}})] \\ & \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\} \\ = \ & \frac{\nu(\mathbf{x})}{nh^p f(\mathbf{x})} \begin{pmatrix} \rho & 0 & \frac{\phi}{h^2}\text{vech}^T\{I\} \\ 0 & \frac{J_2}{\mu_2^2 h^2} I & 0 \\ \frac{\phi}{h^2}\text{vech}\{I\} & 0 & \frac{1}{h^4}E^{-1}E_J E^{-1} - \frac{\mu_2(J_2 - J_0\mu_2)}{(\mu_4 - \mu_2^2)^2 h^4}\text{vech}\{I\}\text{vech}^T\{I\} \end{pmatrix} \\ & + \frac{1}{nh^p}\text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}[O(h) + O_p((nh^p)^{-\frac{1}{2}})]\text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}. \end{aligned}$$

where ρ and ϕ are defined at the end of Theorem 3.2, and constant matrices E and E_J are defined in (3.27) and (3.37), respectively.

This completes the proof of Theorem 3.2. □

Chapter 4

NONPARAMETRIC ESTIMATION WITH TIME SERIES

4.1 Introduction

The multivariate locally weighted polynomial fit with independent observations has been considered in Chapter 3, where the two important cases, the local linear fit and the local quadratic fit, have been studied. In this chapter, the locally weighted polynomial fit is extended to time series. Nonparametric smoothing has been a useful tool for time series analysis, e.g. in model identification and nonlinear prediction, see Tong (1990). Nonparametric estimation with dependent data has often been studied in the literature, see Rosenblatt (1990) and references therein. We will focus on the locally weighted polynomial fit and its applications to estimation of autoregression and its partial derivatives.

Since the observations in time series are dependent, the conditioning approach employed in Chapter 3 is no longer appropriate. Instead, following Masry and Fan

(1993), who have studied the univariate local polynomial fit with time series, we will establish the joint asymptotic normality of the estimators under general conditions. Furthermore, under these conditions, the asymptotic bias and asymptotic covariance matrix are shown to be the same as those in the independent case.

We will assume that the stationary vector time series comes from a general regression-type time series model, typically a nonlinear autoregressive model, and satisfies a short-range dependence condition as defined, e.g. in Castellana and Leadbetter (1986), which is based on the differences between bivariate joint density and product of marginal densities. Under these assumptions, a central limit theorem for martingale arrays can be used to prove the joint asymptotic normality of the estimators. In the context of a general stationary sequence, nonparametric estimation has been considered by Masry and Fan (1993), Rosenblatt (1990), and Castellana and Leadbetter (1986) using stronger mixing conditions.

At last, we discuss some issues in nonparametric fit from a genuinely chaotic time series. A chaotic time series is often finite-dimensional, e.g. the chaotic time series observed from a deterministic system is always so. In general, the time series can have a fractal (nonintegral) dimension. We call a time series *fractal time series*, if its finite-dimensional probability measure has a pointwise fractal dimension. It is noted that a probability measure which has a pointwise dimension can be a singular measure, so it may not have a density.

We will discuss extension of nonparametric estimation methods, in particular the locally weighted polynomial fit to fractal time series. To fix ideas, we consider the Nadaraya-Watson estimator of a nonparametric regression. We assume that the probability measure for the predictor variables has a pointwise fractal dimension. We will establish a convergence rate which involves only the fractal dimension. For the general local polynomial fit, we give a conjecture on the convergence rates for the estimators of regression and partial derivatives.

A related problem which also arises in fitting a chaotic time series is the deterministic fit, or the interpolation of data. See Farmer and Sidorowich (1987, 1988) for more details in connection with nonlinear prediction in chaotic time series. For simplicity, we consider fitting a regression model without the noise term from iid observations. It will be noted that the approximation error is the same as the asymptotic bias of the noisy case. The bandwidth h should be chosen large enough to have sufficient points for the interpolation problem. If the probability measure of predictor variables has pointwise dimension d, it is expected that h scales as $n^{-1/d}$, and so an approximation error can be given accordingly which depends on the interpolation method used.

This chapter is organized as follows. Section 4.2 gives a general regression-type setup for time series. Section 4.3 considers the local linear fit. The main case, the local quadratic fit, is studied in Section 4.4. In Section 4.5, nonparametric estimation in fractal time series is discussed. Section 4.6 discusses the interpolation case. The proofs of Theorem 4.1 and Theorem 4.2 are given in Section 4.7.

4.2 Regression-type Time Series Model

Consider the following regression-type model in time series,

$$Y_i = m(X_i) + \nu^{1/2}(X_i)\varepsilon_i, \;\; i = 1, 2, \dots, \tag{4.1}$$

where Y_i's are scalar response variables and X_i's are $\mathcal{R}^p$-valued predictors. Here, we assume

(A) IID Noises. The noises $\{\varepsilon_i\}$ are iid scalar random variables with zero mean and unit variance. Furthermore, it is assumed that ε_i is independent of $X_i, X_{i-1}, \dots, X_1$.

(B) Ergodicity. Vector sequence $\{X_i\}$ is stationary and ergodic.

(C) Short-range dependence. Under (B), and assume that the joint density of X_1, X_{j+1} exists for any $j \geq 1$, which is denoted by $f_j(\cdot, \cdot)$, and the marginal density is denoted by $f(\cdot)$. Define an index by

$$\beta_n \triangleq \sup_{\mathbf{u},\mathbf{v}\in\mathcal{R}^p} \sum_{j=1}^{n} |f_j(\mathbf{u},\mathbf{v}) - f(\mathbf{u})f(\mathbf{v})|. \tag{4.2}$$

We assume $\beta_n = O(1)$, that is, there exists an $M > 0$ such that $\beta_n \leq M$ for all n.

(D) Moment Condition. There exists a $\delta > 0$ so that $E|\varepsilon|^{2+\delta}$ is finite.

The above general setup includes most models in multivariate and univariate time series. In particular, it includes the nonlinear autoregressive (NAR) model given by

$$x_{i+1} = m(x_i, x_{i-1}, \cdots, x_{i-p+1}) + \nu^{1/2}(x_i, x_{i-1}, \cdots, x_{i-p+1})\varepsilon_{i+1}. \tag{4.3}$$

Usually it is assumed that noises $\{\varepsilon_i\}$ are iid random variables with zero mean and unit variance. Moreover, we assume that ε_1 is independent of initial values $x_0, x_{-1}, \ldots, x_{-p+1}$. Consequently, this implies that ε_i is independent of $x_{i-1}, x_{i-2}, \ldots$ for any $i \geq 1$. To see how NAR fits into the general regression setup, let $Y_i = x_i, X_i = (x_{i-1}, x_{i-2}, \cdots, x_{i-p})^T$, then (4.3) has the form of (4.1).

In our general regression setup, under (A), it is easy to verify that the process $\{Y_i, \mathcal{F}_i^{XY}\}$ is a Markov chain, and the system $\{\varepsilon_i, \mathcal{F}_i^{XY}\}$ satisfies the martingale difference property: $E\{\varepsilon_i|\mathcal{F}_{i-1}^{XY}\} = 0$. Here we define

$$\mathcal{F}_i^{XY} = \sigma\{Y_i, X_{i+1}, Y_{i-1}, X_i, \ldots, Y_2, X_1\} = \sigma\{X_{k+1}, \varepsilon_k : k \leq i\}.$$

Condition (C) is also used by Castellana and Leadbetter (1986), and Rosenblatt (1990). Castellana and Leadbetter (1986) have called β_n in (4.2) a "dependence index". Assumption (A) can be avoided, but at the expense of using some extra mixing conditions, such as those used in Masry and Fan (1993), Castellana and Leadbetter (1986), and Rosenblatt (1990). We feel that our model setup seems more natural than a more general stationary sequence context where stronger mixing conditions are often imposed, which are hard to verify in practice. The following example shows how condition (C) is satisfied in the case of a stationary normal sequence.

Example. We consider the scalar stationary Gaussian process $\{X_i\}$ with mean zero, autocovariance function r_k, we want to show that the process satisfies (C) if r_j is summable, that is, $\sum_{j=0}^{\infty} |r_j| < \infty$.

Proof: Without loss of generality, we can assume that the variance is one. Since $r_k \to 0$ as $k \to \infty$, the joint density function of X_0, X_j evaluated at (x, y), given by

$$g(x,y,r_j) = f_j(x,y) = \frac{1}{2\pi\sqrt{1-r_j^2}} \exp\{-\frac{x^2+y^2-2r_jxy}{2(1-r_j^2)}\},$$

has expansion as $j \to \infty$,

$$g(x,y,r_j) \approx g(x,y,0) + r_j g'(x,y,0),$$

where $g'(x,y,0)$ is $\frac{\partial g(x,y,r)}{\partial r}$ evaluated at $r = 0$. Some calculations shows that

$$\frac{\partial g(x,y,r)}{\partial r} = \frac{1}{2\pi(1-r^2)}\{-\frac{r}{\sqrt{1-r^2}} + (1+r^2)xy - r(x+y)\} \exp\{-\frac{x^2+y^2-2r_jxy}{2(1-r_j^2)}\}.$$

So

$$g'(x,y,0) = \frac{xy}{2\pi} \exp\{-\frac{x^2+y^2}{2}\}.$$

Thus, we have shown that for j large,

$$f_j(x,y) - f(x)f(y) \approx r_j \frac{xy}{2\pi} f(x)f(y),$$

where the right hand side is clearly bounded by $M_0|r_j|$ for some constant M_0 by noting that $xf(x) = \frac{x}{\sqrt{2\pi}}\exp\{-\frac{x^2}{2}\}$ is bounded. Thus,

$$\beta_n = \sup_{x,y \in \mathcal{R}} \sum_{i=1}^{n} |f_j(x,y) - f(x)f(y)| \leq M_0 \sum_{i=1}^{n} |r_j|,$$

and (C) is satisfied if $\sum_{j=0}^{\infty} |r_j| < \infty$. □

A similar statement can be made for a stationary normal vector sequence. Consequences of (C) will be discussed in Lemmas 4.5-4.7, where the connection to the usual sense of short-range dependence in terms of the covariance function will be further clarified.

4.3 Local Linear Fit

In this section, we will study the multivariate local linear fit with dependent data. The local linear fit with independent observations is studied in Section 3.3. As in Section 3.3, we will use the same notations and assume that the kernel K is spherically symmetric and satisfies $\int u_1^8 K(u_1, \cdots, u_p) du_1 \cdots du_p < \infty$. Under model (4.1) and the assumptions (A)-(D), joint asymptotic normality of estimators from the local linear fit is established in the following theorem.

Theorem 4.1 *For l distinct interior points $\mathbf{x}_1, \ldots, \mathbf{x}_l$ inside the support of design density f and $f(\mathbf{x}_j) > 0, \nu(\mathbf{x}_j) > 0$ for all j, and there exist open neighborhoods U_i of $\mathbf{x}_i$ such that $m \in C^3(U_j), f \in C^1(U_j), \nu \in C^0(U), j = 1, 2, \ldots, l$. Then for $h \to 0, nh^p \to \infty$ as $n \to \infty$, the local linear estimators $\widehat{\beta}(\mathbf{x}_1), \ldots, \widehat{\beta}(\mathbf{x}_l)$ in (3.2) are asymptotically independent and jointly normal. In particular, at each point say $\mathbf{x} = (x_1, \cdots, x_p)^T$, we have*

$$(nh^p)^{\frac{1}{2}} \mathrm{diag}\{1, hI\}\{\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) + B(\mathbf{x}, h)\}$$

is asymptotically normal $N(0, \Sigma(\mathbf{x}))$. Here $B(\mathbf{x}, h)$ is the asymptotic bias given by:

$$B(\mathbf{x}, h) = \begin{pmatrix} \frac{1}{2} h^2 \mu_2 \nabla_m^2(\mathbf{x}) + o(h^3) \\ \frac{h^2}{3!\mu_2} b(m, K) + \frac{h^2}{2\mu_2 f(x)} b_1(m, K) + o(h^2) \end{pmatrix}. \tag{4.4}$$

The asymptotic covariance matrix is given by

$$\Sigma(\mathbf{x}) = \begin{pmatrix} \frac{\nu(\mathbf{x}) J_0}{f(\mathbf{x})} & 0 \\ 0 & \frac{\nu(\mathbf{x}) J_2}{{\mu_2}^2 f(\mathbf{x})} I \end{pmatrix}.$$

Here $\nabla_m^2(\mathbf{x}), b(m, K), b_1(m, K), \mu_\ell, J_\ell$ are defined as in Theorem 3.1.

REMARKS ON THEOREM 4.1:

For part of the results on the regression estimator to hold only, the assumptions $m \in C^3(U), f \in C^1(U)$ are not necessary. Instead, weaker assumptions such as $m \in C^2(U), f \in C^0(U)$ will suffice. It can be checked easily that the asymptotic bias and asymptotic covariance matrix have the same forms as if (Y_i, X_i)'s are iid.

4.4 Local Quadratic Fit

We consider the main case of the multivariate local quadratic fit with dependent data. The local quadratic estimators with independent observations are given in Section 3.4. The same notations will be used here, and the kernel K is assumed to be spherically symmetric and satisfies $\int u_1^{12} K(u_1, \cdots, u_p) du_1 \cdots du_p < \infty$. Under model (4.1) and assumptions (A)-(D), the joint asymptotic normality of the local quadratic estimator is given in next theorem.

Theorem 4.2 *For l distinct interior points $\mathbf{x}_1, \ldots, \mathbf{x}_l$ inside the support of design density f and $f(\mathbf{x}_j) > 0, \nu(\mathbf{x}_j) > 0$ for all j, if there exists open neighborhoods U_i of $\mathbf{x}_i$ such that $m \in C^4(U_j), f \in C(U_j), \nu \in C^0(U), j = 1, 2, \ldots, l$, then for $h \rightarrow 0, nh^p \rightarrow \infty$* as *$n \rightarrow \infty$, the local quadratic estimators $\widehat{\beta}(\mathbf{x}_1), \ldots, \widehat{\beta}(\mathbf{x}_l)$ in (3.7) are asymptotically independent and jointly normal. In particular, at each point say $\mathbf{x} = (x_1, \cdots, x_p)^T$, we have*

$$(nh^p)^{\frac{1}{2}} \mathrm{diag}\{1, hI_1, h^2 I_2\}(\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) - B(\mathbf{x}, h))$$

is asymptotically normal $N(0, \Sigma(\mathbf{x}))$. Here I_1, I_2 are identity matrices of dimension p and $p(p+1)/2$, respectively, and $B(\mathbf{x}, h)$ is the asymptotic bias given by:

$$B(\mathbf{x}, h) = \begin{pmatrix} \frac{h^4}{4!}\theta(m, K) + \frac{h^4}{3! f(\mathbf{x})}\theta_1(m, K) + o(h^4) \\ \frac{h^2}{3! \mu_2} b(m, K) + o(h^3) \\ h^2 \gamma(m, K) + \frac{h^2}{f(\mathbf{x})}\gamma_1(m, K) + o(h^2) \end{pmatrix}.$$

The asymptotic covariance matrix is given by:

$$\Sigma(\mathbf{x}) = \begin{pmatrix} \frac{\rho\nu(\mathbf{x})}{f(\mathbf{x})} & 0 & \frac{\phi\nu(\mathbf{x})}{f(\mathbf{x})}\mathrm{vech}^T\{I\} \\ 0 & \frac{J_2 \nu(\mathbf{x})}{\mu_2^2 f(\mathbf{x})} I & 0 \\ \frac{\phi\nu(\mathbf{x})}{f(\mathbf{x})}\mathrm{vech}\{I\} & 0 & \frac{\nu(\mathbf{x})}{f(\mathbf{x})}(\Lambda - \frac{\mu_2 (J_2 - J_0 \mu_2)}{(\mu_4 - \mu_2^2)^2}\mathrm{vech}\{I\}\mathrm{vech}^T\{I\}) \end{pmatrix}. \quad (4.5)$$

Here $b(m, K), \theta(m, K), \gamma(m, K), \gamma_1(m, K), \rho, \phi, \Lambda$ are defined as in Theorem 3.2.

REMARKS ON THEOREM 4.2:

1. For part of the results on the first-order partial derivative, the assumptions $m \in C^4(U), f \in C^1(U)$ are not necessary. Instead, the weaker assumptions $m \in C^3(U), f \in C^0(U)$ will suffice.

2. The assumptions $m \in C^4(U), f \in C^1(U)$ are necessary for the regression estimator and the estimators of second-order (mixed) partial derivatives, i.e. Hessian matrix estimation. Under the weaker assumption $m \in C^3(U), f \in C^0(U)$, their bias have lower orders of $O(h^3)$ and $O(h)$, respectively.

3. The asymptotic bias and asymptotic covariance matrix are the same as if (Y_i, X_i)'s are iid.

In next two sections, we will discuss some issues in nonparametric fit from a chaotic time series. The density assumption on the predictor variables will be relaxed, and the multivariate predictor can have a singular probability measure (which means the measure is continuous and is singular with respect to the Lebesgue measure) . Another issue is deterministic fit or interpolation. The implications of the present results for the approximation error in the interpolation problem will be pointed out.

4.5 Nonparametric Estimation with Fractal Time Series

In this section, we will discuss some open questions in nonparametric estimation from a genuinely chaotic time series. A genuinely chaotic time series is often finite-dimensional (up to certain scale) and has a fractal (nonintegral) dimension. For example, a time series observed from a deterministic system is always finite-dimensional. A simple example is given by $x_{t+1} = 4x_t(1-x_t)$, with x_0 randomly distributed according to density $f(x) = 1/(\pi\sqrt{x(1-x)})$. This sequence is stationary and the dimension of the embedded time series $(x_t, x_{t-1}, \cdots, x_{t-p+1})^T$ is always one for any $p \geq 1$ using any dimension definition. More examples are given in Chapter 2. In this section, we discuss nonparametric estimation in a noisy time series. The nonparametric fit in a deterministic time series or the interpolation problem will be discussed in the next section.

Specifically, we will consider a fractal time series, that is, we assume that the finite-dimensional probability measure for $(x_t, x_{t-1}, \ldots, x_{t-p+1})^T$ has a fractal pointwise dimension d(for large enough p). A measure with a pointwise dimension may be continuous and singular with respect to the Lebesgue measure and does not have a density function in the usual sense. For fractal time series , we expect that the convergence rate of a nonparametric estimator is independent of p and it depends on the fractal dimension d only. So the "curse of dimension" problem associated with fitting high-dimensional models may not occur with fractal time series. For simplicity, we will assume iid observations in a regression setup. We expect that similar results may still hold for time series under some short-range dependence conditions, but we will leave those as future problems.

Nonparametric regression. Consider model (4.1), where for simplicity we assume that $\{(Y_i, X_i), i = 1, 2, \ldots, n\}$ are iid observations, and $\nu(\mathbf{x}) = \sigma^2$. We assume that the probability measure of X_1 denoted by ρ has a pointwise fractal dimension d, which will be typically smaller than p. See Chapter 2 for more discussions on singular measures and fractal dimensions.

For simplicity we will make the following stronger assumption, letting $B_r(\mathbf{x})$ denote a sphere of radius r centered at $\mathbf{x}$, then for ρ-almost all $\mathbf{x}$,

$$\rho(B_r(\mathbf{x})) = c(\mathbf{x}) r^d (1 + o(1)) \text{ as } r \to 0, \tag{4.6}$$

or as a shorthand notation,

$$\rho(B_r(\mathbf{x})) \sim c(\mathbf{x}) r^d, \text{ as } r \to 0,$$

where we use "$\sim$" to mean that the ratio of both sides converges to 1. If ρ is absolutely continuous, $c(\mathbf{x})$ coincides with the density function $f(\mathbf{x})$ (apart from a normalizing constant).

Note that we can rewrite the above as

$$EK(\frac{X-\mathbf{x}}{r}) \sim c(\mathbf{x})r^d, \text{ as } r \to 0,$$

where $K(\mathbf{x}) = 1_{\{\|\mathbf{x}\| \le 1\}}$. We can extend the scaling relation to a general spherically symmetric function $K(\mathbf{x}) = k(\|\mathbf{x}\|)$, as shown in the following lemma.

Lemma 4.1 *Suppose* $\mathbf{x}$ *is such that (4.6) holds. Assume that* k *is a univariate kernel function with finite support, i.e.* $k(x) = 0$ *for* x *outside* $[0,1]$*, and satisfies the Lipschitz condition, i.e. there exists* c_k*,* $0 < \alpha \le 1$ *such that*

$$|k(x) - k(y)| \le c_k |x-y|^\alpha, \tag{4.7}$$

for all x, y *in* $[0,1]$*. Then,*

$$Ek(\frac{\|X-\mathbf{x}\|}{h}) \sim c(\mathbf{x})h^d d \int k(y)y^{d-1}dy, \text{ as } h \to 0. \tag{4.8}$$

PROOF: Given any partition on $[0,1]$:

$$0 = a_0 < a_1 < a_2 < \cdots a_{n-1} < a_n = 1, \text{ and let } \Delta = \max_i(a_{i+1} - a_i).$$

Write

$$\begin{aligned} k(y) &= \sum_{i=0}^{n-1} k(a_i) 1_{\{a_i < y < a_{i+1}\}} + \sum_{i=0}^{n-1} (k(y) - k(a_i)) 1_{\{a_i < y < a_{i+1}\}} \\ &\triangleq I(y) + II(y), \end{aligned}$$

and

$$Ek(\frac{\|X-\mathbf{x}\|}{h}) = EI(\frac{\|X-\mathbf{x}\|}{h}) + EII(\frac{\|X-\mathbf{x}\|}{h}). \tag{4.9}$$

The first term in (4.9) is given by

$$\begin{aligned} EI(\frac{\|X-\mathbf{x}\|}{h}) &= \sum_{i=0}^{n-1} k(a_i) E1_{\{a_i < \frac{\|X-\mathbf{x}\|}{h} < a_{i+1}\}} \\ &= \sum_{i=0}^{n-1} k(a_i)(\rho(B_{ha_{i+1}}(\mathbf{x}) - \rho(B_{ha_i}(\mathbf{x}))) \\ &= \sum_{i=0}^{n-1} k(a_i) c(\mathbf{x}) h^d (a_{i+1}^d - a_i^d)(1 + o(1)) \\ &= c(\mathbf{x}) h^d \{\sum_{i=0}^{n-1} k(a_i)(a_{i+1}^d - a_i^d)\}(1 + o(1)) \\ &\quad \text{as } h \to 0. \end{aligned}$$

The boundedness of k is used in the last operation.

Since Δ is arbitrary, by letting $\Delta \to 0$, the term inside $\{\}$ tends to

$$\int k(y)dy^d = d\int k(y)y^{d-1}dy,$$

or write as

$$\sum_{i=0}^{n-1} k(a_i)(a_{i+1}^d - a_i^d) = d\int k(y)y^{d-1}dy + o_\Delta(1).$$

Now consider the 2nd term in (4.9),

$$\begin{aligned} |EII(\frac{\|X-\mathbf{x}\|}{h})| &\leq \sum_{i=0}^{n-1} E|k(\frac{\|X-\mathbf{x}\|}{h}) - k(a_i)|1_{\{a_i < \frac{\|X-\mathbf{x}\|}{h} < a_{i+1}\}} \\ &\leq c_k\Delta^\alpha \sum_{i=0}^{n-1} \rho\{ha_i < \|X-\mathbf{x}\| < ha_{i+1}\} \\ &= c_k\Delta^\alpha \rho\{\|X-\mathbf{x}\| < h\} \\ &= c_k\Delta^\alpha c(\mathbf{x})h^d(1+o(1)) \\ &\quad \text{as } h \to 0. \end{aligned}$$

Thus, (4.9) becomes

$$\begin{aligned} &Ek(\frac{\|X-\mathbf{x}\|}{h}) = \\ &\quad c(\mathbf{x})h^d\{d\int k(y)y^{d-1}dy + o_\Delta(1)\}(1+o(1)) + O(\Delta^\alpha h^\alpha)(1+o(1)). \end{aligned}$$

Since Δ is arbitrary, by letting $\Delta \to 0$, and noticing that the second term is negligible compared with the order $O(h^d)$ of the first term, we have

$$Ek(\frac{\|X-\mathbf{x}\|}{h}) = c(\mathbf{x})h^d\{d\int k(y)y^{d-1}dy\}(1+o(1)).$$

The lemma is proved. □

Next, we will prove a theorem on the Nadaraya-Watson estimator given by

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^n K(\frac{X_i - \mathbf{x}}{h})Y_i}{\sum_{i=1}^n K(\frac{X_i - \mathbf{x}}{h})}.$$

We assume that $K(\mathbf{x}) = k(\|\mathbf{x}\|$ with k satisfying the conditions in Lemma 4.1. We also assume that $E|\varepsilon_1|^{2+\delta} < \infty$ for some $\delta > 0$.

Theorem 4.3 *Assume that probability ρ of X_1 has the scaling relation (4.6). For any given $\mathbf{x}$ such that (4.6) holds with $c(\mathbf{x}) > 0$. Let $U_{\mathbf{x}}$ be an open neighborhood of $\mathbf{x}$, and assume that m is Lipschitz continuous with exponent s, i.e. there exists a $c_m > 0$, such that for any $\mathbf{y}_1, \mathbf{y}_2$ in $U_{\mathbf{x}}$,*

$$|m(\mathbf{y}_1) - m(\mathbf{y}_2)| \leq c_m \|\mathbf{y}_1 - \mathbf{y}_2\|^s, \text{ where } 0 < s \leq 1.$$

Then as $h \to 0, nh^d \to \infty$,

$$\sqrt{nh^d}\{\hat{m}(\mathbf{x}) - m(\mathbf{x}) - b_n\}$$

is asymptotically normal $N(0, \delta^2)$, where constant $b_n = O(h^s)$, and

$$\delta^2 = \frac{\sigma^2 \int k(y)^2 y^{d-1} dy}{(\int k(y) y^{d-1} dy)^2 dc(\mathbf{x})}.$$

By choice of optimal $h = O(n^{-\frac{1}{d+2s}})$, the pointwise convergence rate (in probability) of $\hat{m}(x)$ is seen to be $O_p(n^{-\frac{s}{d+2s}})$.

Proof: Note that

$$\begin{aligned} \hat{m}(\mathbf{x}) - m(\mathbf{x}) &= \frac{\sum_{i=1}^n (m(X_i) - m(\mathbf{x}))K(\frac{X_i - \mathbf{x}}{h})}{\sum_{i=1}^n K(\frac{X_i - \mathbf{x}}{h})} + \frac{\sum_{i=1}^n \sigma\varepsilon_i K(\frac{X_i - \mathbf{x}}{h})}{\sum_{i=1}^n K(\frac{X_i - \mathbf{x}}{h})} \\ &\triangleq B_n + R_n. \end{aligned}$$

Now let's first consider the 2nd term R_n. By Lemma 4.1,

$$EK(\frac{X_1 - \mathbf{x}}{h}) \sim dc(\mathbf{x})h^d \int k(y) y^{d-1} dy, \tag{4.10}$$

$$\text{Var}\{K(\frac{X_1 - \mathbf{x}}{h})\} \sim dc(\mathbf{x})h^d \int k(y)^2 y^{d-1} dy.$$

Using

$$\sum_{i=1}^n K(\frac{X_i - \mathbf{x}}{h}) = nEK(\frac{X_1 - \mathbf{x}}{h}) + O_p(\sqrt{n\text{Var}\{K(\frac{X_1 - \mathbf{x}}{h})\}}),$$

by Proposition 3.2 in Chapter 3, we have as $h \to 0, nh^p \to \infty$,

$$S_n = \frac{1}{nh^d} \sum_{i=1}^n K(\frac{X_i - \mathbf{x}}{h}) = h^{-d} EK(\frac{X_1 - \mathbf{x}}{h}) + O_p((nh^d)^{-\frac{1}{2}}). \tag{4.11}$$

Denote the numerator of R_n by

$$T_n \triangleq \sum_{i=1}^{n} \sigma \varepsilon_i K(\frac{X_i - \mathbf{x}}{h}). \tag{4.12}$$

Note that $ET_n = 0$,

$$\mathrm{Var} T_n = n\sigma^2 \mathrm{Var}\{K(\frac{X_i - \mathbf{x}}{h})\} \sim nh^d \sigma^2 dc(\mathbf{x}) \int k(y)^2 y^{d-1} dy.$$

And

$$\begin{aligned} \Lambda_n &\triangleq \sum_{i=1}^{n} E|\sigma\varepsilon_i K(\frac{X_i - \mathbf{x}}{h})|^{2+\delta} \\ &= n\sigma^{2+\delta} E|\varepsilon_1|^{2+\delta} EK(\frac{X_i - \mathbf{x}}{h})^{2+\delta} \\ &\sim nh^d \sigma^{2+\delta} dc(\mathbf{x}) E|\varepsilon_1|^{2+\delta} \int k^{2+\delta}(y) y^{d-1} dy. \end{aligned}$$

Thus,

$$\frac{\Lambda_n}{(\mathrm{Var} T_n)^{\frac{2+\delta}{2}}} = O((nh^d)^{-\delta/2}) \tag{4.13}$$

which tends to zero as $nh^d \to \infty$. This implies that conditions of the central limit theorem for sums of triangular arrays with independence within rows (e.g. P. 32, Corollary 1.9.3 in Serfling (1981)) are satisfied.

We have proved,

$$\frac{1}{\sqrt{nh^d}} T_n \xrightarrow{d} N(0, \delta_1^2), \text{ where } \delta_1^2 = \sigma^2 dc(\mathbf{x}) \textstyle\int k(y)^2 y^{d-1} dy.$$

Combined with (4.11) and (4.10), we have

$$\begin{aligned} \sqrt{nh^d} R_n &= \frac{(\sqrt{nh^d})^{-1} T_n}{S_n} \\ &\xrightarrow{d} N(0, \delta^2), \text{ where } \delta^2 = \tfrac{\sigma^2 \int k(y)^2 y^{d-1} dy}{(\int k(y) y^{d-1} dy)^2 dc(\mathbf{x})}. \end{aligned}$$

That is, we obtain

$$\sqrt{nh^d}\{\hat{m}(\mathbf{x}) - m(\mathbf{x}) - B_n\} \xrightarrow{d} N(0, \delta^2). \tag{4.14}$$

Now consider the first term B_n. Since B_n is random, the asymptotic normality result in (4.14) cannot be used directly in practice. We will show next how to replace B_n by a constant.

Applying Lemma 4.1 and using the Lipschitz continuity of m and finite support of K, we have

$$\begin{aligned} |E\{m(X_1)-m(\mathbf{x})\}K(\frac{X_1-\mathbf{x}}{h})| &\leq E|m(X_1)-m(\mathbf{x})|K(\frac{X_1-\mathbf{x}}{h}) \\ &\leq c_m E\|X_1-\mathbf{x}\|^s K(\frac{X_1-\mathbf{x}}{h}) \\ &\sim c_m dc(\mathbf{x})h^{d+s}\int y^s k(y) y^{d-1}dy. \end{aligned}$$

I.e.

$$E\{m(X_1)-m(\mathbf{x})\}K(\frac{X_1-\mathbf{x}}{h}) = O(h^{d+s}) \tag{4.15}$$

Similarly,

$$E\{m(X_1)-m(\mathbf{x})\}^2K^2(\frac{X_1-\mathbf{x}}{h}) = O(h^{d+2s}).$$

So

$$\mathrm{Var}\{(m(X_1)-m(\mathbf{x}))K(\frac{X_1-\mathbf{x}}{h})\} = O(h^{d+2s}).$$

Applying above to

$$\begin{aligned} &\sum_{i=1}^{n}(m(X_i)-m(\mathbf{x}))K(\frac{X_i-\mathbf{x}}{h}) = \\ &\qquad nE\{m(X_1)-m(\mathbf{x})\}K(\frac{X_1-\mathbf{x}}{h}) + O_p(\sqrt{n\mathrm{Var}\{(m(X_1)-m(\mathbf{x}))K(\frac{X_1-\mathbf{x}}{h})\}}), \end{aligned}$$

which is given by using Proposition 3.2, we obtain

$$\begin{aligned} &\frac{1}{nh^d}\sum_{i=1}^{n}(m(X_i)-m(\mathbf{x}))K(\frac{X_i-\mathbf{x}}{h}) = \\ &\qquad h^{-d}E\{m(X_1)-m(\mathbf{x})\}K(\frac{X_1-\mathbf{x}}{h}) + h^s O_p((nh^d)^{-1/2}). \end{aligned}$$

Define

$$b_n = \frac{h^{-d}E\{m(X_1)-m(\mathbf{x})\}K(\frac{X_1-\mathbf{x}}{h})}{h^{-d}EK(\frac{X_1-\mathbf{x}}{h})} \tag{4.16}$$

which is of order $O(h^s)$ from (4.10) and (4.15).

Combining (4.11) and (4.16), we have

$$B_n = b_n + h^s O_p((nh^d)^{-1/2}), \text{ as } h \to 0, nh^p \to \infty.$$

So replacing B_n by constant b_n in (4.14) will not affect the asymptotic normality. The theorem is thus proved. □

A general conjecture. Note that the generalized scaling relation (4.8) for the expectation of a spherically symmetric smooth kernel function is obtained from the assumption (4.6) on the probability of a small ball. The question arises whether the scaling relation (4.8) still holds for more general functions. In particular, in order to extend Theorem 4.3 to a more general case, such as estimating regression and partial derivatives by fitting a higher-order local polynomial, we need similar scaling relation to (4.8) to hold for moments of a spherically symmetric smooth function. We make the following conjecture. There is a possibility that more conditions on ρ may be needed.

Conjecture 4.1 *Assume that a probability measure ρ satisfies (4.6). For a smooth spherically symmetric function K, we have*

$$E(\frac{X_{11} - x_1}{h})^{l_1} \cdots (\frac{X_{1p} - x_p}{h})^{l_p} K(\frac{X - \mathbf{x}}{h}) \sim c(\mathbf{x}) h^d J(K), \tag{4.17}$$

as $h \to 0$, where $X = (X_{11}, \cdots, X_{1p})^T, \mathbf{x} = (x_1, \cdots, x_p)^T$, and $l_1, \ldots, l_p$ are nonnegative integers, and $J(K)$ is a constant depending on K and its moments of order $l_1, \ldots, l_p$.

Under Conjecture 4.1, we can establish similar convergence rates on the regression and partial derivative estimators for the local linear fit and the local quadratic fit. More generally, we give the following general conjecture on the convergence rates for estimators of a nonparametric regression and its partial derivatives of any order from a local polynomial fit.

Let U denote an open neighborhood of a given $\mathbf{x}$. Let C^s (where $s > 0$) be the class of $s-$times continuously differentiable functions if s is an integer, or if s is a noninteger the $[s]$th-order partial derivatives satisfy the Lipschitz continuity with exponent $s-[s]$. Here $[s]$ denotes the integer part of s.

Conjecture 4.2 *For a given* $\mathbf{x}$ *such that (4.6) holds with* $c(\mathbf{x}) > 0$*, and suppose Conjecture 4.1 holds. Assume that* $m \in C^s$*. Then the convergence rate for the estimator of partial derivative of order* $l < s$*(*m *corresponding to* $l = 0$*) from the local polynomial fit of order* k *(where* $k = s - 1$ *if* s *is an integer, and* $k = [s]$ *if* s *is a noninteger) is given by:*

$$O(h^{s-l}) + O_p((nh^{d+2l})^{-\frac{1}{2}}),$$

where h *is the bandwidth. By choosing* $h = O(n^{-\frac{1}{d+2s}})$*, the pointwise convergence rate is given by* $O_p(n^{-\frac{s-l}{d+2s}})$*.*

4.6 Deterministic Fit

A related problem is the deterministic fit or interpolation in time series. More discussions are given in Farmer and Sidorowich (1987, 1988), who have used nonparametric methods for nonlinear prediction in chaotic time series. The interpolation problem is also much studied in the approximation function literature. For simplicity, we consider iid observations (Y_i, X_i)'s which satisfy $Y_i = m(X_i), i = 1, 2, \ldots, n$. We are interested in approximating m or its partial derivatives from data.

The approximation error will depend on the interpolation method used. We want to point out that the approximation error is the same as the asymptotic bias of the noisy case. In the interpolation case, the bandwidth h in the kernel method should be chosen as small as possible to minimize approximation error or bias. However, the bandwidth h should be chosen large enough to have sufficient points for the interpolation problem. If the probability measure X has pointwise dimension d, it is expected that h scales as $n^{-1/d}$, and so an approximation error can be given accordingly. Another way to see this is to notice that the asymptotic variance of the interpolated value is given by the higher-order terms in that of the noisy case. E.g.

in the case of using local linear fit to interpolate m at a point $\mathbf{x}$, the asymptotic MSE is given by

$$O(h^4) + O(h^4)O((nh^d)^{-1}),$$

So choosing $h = O(n^{-1/d})$ will give the minimum convergence rate $n^{-2/d}$.

4.7 Proofs

Proof of Theorem 4.1 is considered to be easier and follows along the lines of Theorem 4.2. So we only prove the latter here.

Some preliminaries. Note that

$$\begin{aligned}\hat{\beta} - \beta &= (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W (Y - \mathbf{X}\beta) \\ &= (\frac{1}{nh^p}\mathbf{X}^T W \mathbf{X})^{-1} \{\frac{1}{nh^p}\mathbf{X}^T W (M - \mathbf{X}\beta)\} \\ &\quad + (\frac{1}{nh^p}\mathbf{X}^T W \mathbf{X})^{-1} \{\frac{1}{nh^p}\mathbf{X}^T W V^{\frac{1}{2}} E\},\end{aligned} \tag{4.18}$$

where

$$M = (m(X_1), \cdots, m(X_n))^T, V = \mathrm{diag}\{\nu(X_1), \cdots, \nu(X_n)\},$$

$$E = (\varepsilon_1, \cdots, \varepsilon_n)^T.$$

We use the same notations S_n, R_n as defined in in (3.13) and (3.14) of Section 3.5.1, respectively. We also write

$$\frac{1}{nh^p}\mathbf{X}^T W V^{\frac{1}{2}} E = \frac{1}{\sqrt{nh^p}}\mathrm{diag}\{1, hI_1, h^2 I_2\} Z_n,$$

where we denote

$$Z_n = \frac{1}{\sqrt{nh^p}} \sum_{i=1}^{n} Z(n,i),$$

where

$$Z(n,i) = \begin{pmatrix} K(\frac{X_i - \mathbf{x}}{h}) \nu^{1/2}(X_i) \varepsilon_i \\ (\frac{X_i - \mathbf{x}}{h}) K(\frac{X_i - \mathbf{x}}{h}) \nu^{1/2}(X_i) \varepsilon_i \\ vech\{(\frac{X_i - \mathbf{x}}{h})(\frac{X_i - \mathbf{x}}{h})^T\} K(\frac{X_i - \mathbf{x}}{h}) \nu^{1/2}(X_i) \varepsilon_i \end{pmatrix}.$$

Using these notations we write

$$\hat{\beta} - \beta = \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}\{S_n^{-1}R_n + (nh^p)^{-\frac{1}{2}}S_n^{-1}Z_n\}. \tag{4.19}$$

Note that the argument of n in $S(n,i), R(n,i), Z(n,i)$ denotes the dependence on sample size n through h, which tends to zero proportionally to n. We need next several lemmas in the proof of Theorem 4.2.

Several lemmas. The idea of establishing joint normality of $\hat{\beta}$ is based on (4.19), consisting of three lemmas, which assume model (4.1) and (A)-(D).

Lemma 4.2 *As $nh^p \to \infty$,*

$$S_n = A(h) + O_p((nh^p)^{-\frac{1}{2}}), \tag{4.20}$$

where $A(h)$ is defined in Lemma 3.1.

Lemma 4.3 *Assume $m \in C^4(U), f \in C^1(U)$, as $h \to 0, nh^p \to \infty$,*

$$R_n = h^3\{R(h,\mathbf{x}) + o(h) + O_p((nh^p)^{-\frac{1}{2}})\}, \tag{4.21}$$

where $R(h,\mathbf{x})$ is defined in Lemma 3.5 in Chapter 3.

Lemma 4.4

$$Z_n \to N(0, \Sigma_1), \tag{4.22}$$

where

$$\Sigma_1 = v(\mathbf{x})f(\mathbf{x}) \int \begin{pmatrix} 1 \\ \mathbf{u} \\ \text{vech}\{\mathbf{u}\mathbf{u}^T\} \end{pmatrix} (1, \mathbf{u}^T, \text{vech}^T\{\mathbf{u}\mathbf{u}^T\}) K(\mathbf{u})^2 d\mathbf{u} + O(h).$$

Assuming Lemmas 4.2 to 4.4, and Lemma 4.7 (to be given later), we prove Theorem 4.2.

Proof of Theorem 4.2.

From (4.19), we have

$$(nh^p)^{\frac{1}{2}}\text{diag}\{1, hI_1, h^2I_2\}\{\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) - B_n\} = S_n^{-1}Z_n, \tag{4.23}$$

where

$$B_n = \text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}S_n^{-1}R_n.$$

By Lemma 4.2, Lemma 4.4, and Lemma 3.4, we have that the right-hand side of (4.23) tends in distribution to $N(0, \Sigma)$, where

$$\Sigma = A^{-1}(h)\Sigma_1 A^{-1}(h),$$

which has the expansion as given in the Theorem by using also the calculations of Lemma 3.6 of Chapter 3.

Note that B_n in (4.23) is random. Next we need to show how B_n can be replaced by a constant vector. By Lemma 4.2 and Lemma 4.3, and Lemma 3.4,

$$B_n = h^3\text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}\{A^{-1}(h)R(h, \mathbf{x}) + o(h) + O_p(\{nh^p\}^{-1/2})\},$$

where $A^{-1}(h)$ is given in Lemma 3.3. Defining

$$B(\mathbf{x}, h) = h^3\text{diag}\{1, h^{-1}I_1, h^{-2}I_2\}\{A^{-1}(h)R(h, \mathbf{x}) + o(h)\}, \tag{4.24}$$

it is seen that

$$B_n = B(\mathbf{x}, h) + h^3 O_p(\{nh^p\}^{-1/2}).$$

That is, as $h \to 0, nh^p \to \infty$, replacement of B_n by $B(\mathbf{x}, h)$ in (4.23) will not affect the asymptotic normality.

Going through the same calculations as in the proof of bias in Theorem 3.2 of Chapter 3, we have the asymptotic expansions for B_n as given in the Theorem.

The asymptotic independence of the estimator at different points is easily seen from Lemma 4.7 to be stated below. Theorem 4.2 is proved. □

Before we can prove Lemmas 4.2-4.4, we need some results on the consequences of the short-range dependence condition (C).

Short-range dependence. Short-range dependence condition (C) has some important consequences. The first lemma shows the connection of condition (C) to the usual short-range dependence in time series.

Lemma 4.5 *Given any measurable function $g : \mathcal{R}^p \to \mathcal{R}$ such that*

$$Eg(X_1)^2 < \infty, \int |g(\mathbf{u})| d\mathbf{u} < \infty,$$

we have

$$\sum_{i=1}^{\infty} |\mathrm{Cov}(g(X_1), g(X_{i+1}))| < \infty,$$

and consequently,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = Eg(X_1) + O_p(\frac{1}{\sqrt{n}}).$$

Proof: Note that

$$\begin{aligned}\mathrm{Cov}(g(X_0), g(X_i)) &= \int g(\mathbf{u})g(\mathbf{v})f_i(\mathbf{u},\mathbf{v})d\mathbf{u}d\mathbf{v} - \int g(\mathbf{u})g(\mathbf{v})f(\mathbf{u})f(\mathbf{v})d\mathbf{u}d\mathbf{v} \\ &= \int g(\mathbf{u})g(\mathbf{v})(f_i(\mathbf{u},\mathbf{v}) - f(\mathbf{u})f(\mathbf{v}))d\mathbf{u}d\mathbf{v}.\end{aligned}$$

So

$$\begin{aligned}&\sum_{i=1}^{n} \mathrm{Cov}(g(X_1), g(X_{i+1})) \\ &= \int g(\mathbf{u})g(\mathbf{v})\sum_{i=1}^{n}(f_i(\mathbf{u},\mathbf{v}) - f(\mathbf{u})f(\mathbf{v}))d\mathbf{u}d\mathbf{v},\end{aligned}$$

which is bounded by $|\beta_n|(\int |g(\mathbf{u})|d\mathbf{u})^2$. By condition (C), $\beta_n \le M < \infty$ for all n, and so

$$\sum_{i=1}^{\infty} |\mathrm{Cov}(g(X_1), g(X_{i+1}))| \le M(\int |g(\mathbf{u})|d\mathbf{u})^2 < \infty.$$

The rest of the lemma follows from the fact that

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = Eg(X_0) + O_p(\frac{1}{\sqrt{n}}\{\mathrm{Var}(\sum_{i=1}^{n} g(X_i))\}^{\frac{1}{2}}),$$

by proposition 3.2 of Chapter 3. □

Lemma 4.5 shows that condition (C) implies that the covariance function is summable for any properly-defined instantaneous transformation. Thus, such transformed time series is short-range dependent in the usual sense used in time series analysis. The following lemma is needed.

Lemma 4.6 *Under condition (C), for any measurable function g defined on $\mathcal{R}^p$ such that $\int |g(\mathbf{u})|d\mathbf{u} < \infty$, we have as $h \to 0, nh^p \to \infty$,*

$$\mathrm{Var}(\sum_{i=1}^n g(\frac{X_i - \mathbf{x}}{h})) = nh^p f(\mathbf{x}) \int g(\mathbf{u})^2 d\mathbf{u} + O(nh^{2p}), \tag{4.25}$$

and

$$\frac{1}{nh^p} \sum_{i=1}^n g(\frac{X_i - \mathbf{x}}{h}) = f(\mathbf{x}) \int g(\mathbf{u}) d\mathbf{u} + O_p((nh^p)^{-\frac{1}{2}}).$$

Proof:

$$\mathrm{Var}(\sum_{i=1}^n g(\frac{X_i - \mathbf{x}}{h})) = n\mathrm{Var}(g(\frac{X_i - \mathbf{x}}{h})) + \sum_{i=1}^{n-1} (n-i)\mathrm{Cov}(g(\frac{X_{i+1} - \mathbf{x}}{h}), g(\frac{X_1 - \mathbf{x}}{h})).$$

And,

$$\sum_{i=1}^{n-1} (n-i)\mathrm{Cov}(g(\frac{X_{i+1} - \mathbf{x}}{h}), g(\frac{X_1 - \mathbf{x}}{h})) =$$
$$\sum_{i=1}^{n} (n-i) \int g(\frac{\mathbf{u} - \mathbf{x}}{h}) g(\frac{\mathbf{v} - \mathbf{x}}{h})(f_i(\mathbf{u}, \mathbf{v}) - f(\mathbf{u})f(\mathbf{v}))d\mathbf{u}d\mathbf{v}$$

which is bounded by

$$\begin{aligned} & n \int g(\frac{\mathbf{u} - \mathbf{x}}{h}) g(\frac{\mathbf{v} - \mathbf{x}}{h}) \sum_{i=1}^n |f_i(\mathbf{u}, \mathbf{v}) - f(\mathbf{u})f(\mathbf{v})| d\mathbf{u}d\mathbf{v} \\ \leq \; & nM(\int |g(\frac{\mathbf{u} - \mathbf{x}}{h})| d\mathbf{u})^2 \\ = \; & nMh^{2p}(\int |g(\mathbf{u})| d\mathbf{u})^2, \end{aligned}$$

using condition (C).

So as $h \to 0, n \to \infty$,

$$\mathrm{Var}(\sum_{i=1}^n g(\frac{X_i - \mathbf{x}}{h})) = nh^p f(\mathbf{x}) \int g(\mathbf{u})^2 d\mathbf{u} + O(nh^{2p}) = O(nh^p).$$

Note that

$$\sum_{i=1}^{n} g(\frac{X_i - \mathbf{x}}{h}) = nEg(\frac{X_1 - \mathbf{x}}{h}) + O_p(\{\mathrm{Var}(\sum_{i=1}^{n} g(\frac{X_i - \mathbf{x}}{h}))\}^{\frac{1}{2}}), \tag{4.26}$$

by Proposition 3.2 in Chapter 3, we have

$$\sum_{i=1}^{n} g(\frac{X_i - \mathbf{x}}{h}) = nh^p f(\mathbf{x}) \int g(\mathbf{u}) d\mathbf{u} + O_p((nh^p)^{\frac{1}{2}}),$$

i.e.

$$\frac{1}{nh^p} \sum_{i=1}^{n} g(\frac{X_i - \mathbf{x}}{h}) = f(\mathbf{x}) \int g(\mathbf{u}) d\mathbf{u} + O_p((nh^p)^{-\frac{1}{2}}).$$

The lemma is proved. □

Lemma 4.6 implies that $\sum_{i=1}^{n} g(\frac{X_i - \mathbf{x}}{h})$ behaves asymptotically in the same way as if $\{X_i\}$ are independent. The next lemma is used to establish asymptotic independence of the estimators at different points in proof of Theorem 4.2.

Let $G_n(\mathbf{x}) = (nh^p)^{-1/2} \sum_{i=1}^{n} g(\frac{X_i - \mathbf{x}}{h})$.

Lemma 4.7 *Under condition (C), for any* $\mathbf{x} \neq \mathbf{y}, f(\mathbf{x}), f(\mathbf{y}) > 0$, *and any bounded continuous function* g *defined on* R^p *with* $\int |g(\mathbf{u})| d\mathbf{u} < \infty$, *we have as* $h \to 0, nh^p \to \infty$,

$$\mathrm{Cov}\{G_n(\mathbf{x}), G_n(\mathbf{y})\} = o(1).$$

Proof: Note that

$$\begin{aligned} \mathrm{Cov}\{G_n(\mathbf{x}), G_n(\mathbf{y})\} &= h^{-p} \mathrm{Cov}\{g(\frac{X_1 - \mathbf{x}}{h}), g(\frac{X_1 - \mathbf{y}}{h})\} \\ &\quad + (nh^p)^{-1} \sum_{i=1}^{n-1} (n-i) \mathrm{Cov}\{g(\frac{X_{i+1} - \mathbf{x}}{h}), g(\frac{X_1 - \mathbf{y}}{h})\}. \end{aligned}$$

Note that

$$\sum_{i=1}^{n-1} (n-i) \mathrm{Cov}\{g(\frac{X_{i+1} - \mathbf{x}}{h}), g(\frac{X_1 - \mathbf{y}}{h})\} =$$
$$\sum_{i=1}^{n-1} (n-i) \int g(\frac{\mathbf{u} - \mathbf{x}}{h}) g(\frac{\mathbf{v} - \mathbf{y}}{h} (f_j(\mathbf{u}, \mathbf{v}) - f(\mathbf{u}) f(\mathbf{v})) d\mathbf{u} d\mathbf{v},$$

which is bounded by

$$\begin{aligned} & n\int |g(\frac{\mathbf{u}-\mathbf{x}}{h})||g(\frac{\mathbf{v}-\mathbf{y}}{h}| \sum_{i=1}^{n-1}(f_j(\mathbf{u},\mathbf{v})-f(\mathbf{u})f(\mathbf{v}))|d\mathbf{u}d\mathbf{v} \\ & \leq nM(\int |g(\frac{\mathbf{u}-\mathbf{x}}{h})|d\mathbf{u})^2 \\ & = nMh^{2p}\int |g(\mathbf{u})|d\mathbf{u})^2. \end{aligned}$$

Furthermore,

$$\begin{aligned} Eg(\frac{X_1-\mathbf{y}}{h})g(\frac{X_1-\mathbf{x}}{h}) &= \int g(\frac{\mathbf{x}_1-\mathbf{y}}{h})g(\frac{\mathbf{x}_1-\mathbf{x}}{h})f(\mathbf{x}_1)d\mathbf{x}_1 \\ &= h^p\int g(\frac{\mathbf{x}-\mathbf{y}}{h}+\mathbf{u})g(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u}. \end{aligned}$$

As $h \to 0$, it is expected that

$$\int g(\frac{\mathbf{x}-\mathbf{y}}{h}+\mathbf{u})g(\mathbf{u})f(\mathbf{x}+h\mathbf{u})d\mathbf{u} \approx f(\mathbf{x})\int g(\frac{\mathbf{x}-\mathbf{y}}{h}+\mathbf{u})g(\mathbf{u})d\mathbf{u}$$

which is bounded by

$$f(\mathbf{x})\{\int g(\frac{\mathbf{x}-\mathbf{y}}{h}+\mathbf{u})^2d\mathbf{u}\}^{1/2}\{\int g(\mathbf{u})^2d\mathbf{u}\}^{1/2},$$

by the Cauchy-Schwarz inequality. Since $g(\frac{\mathbf{x}-\mathbf{y}}{h}+\mathbf{u}) \to 0$ as $h \to 0$, and g is bounded, by Lebesgue's dominated convergence theorem,

$$\int g(\frac{\mathbf{x}-\mathbf{y}}{h}+\mathbf{u})^2d\mathbf{u} \to 0, \text{ as } h \to 0.$$

Thus,

$$\text{Cov}\{g(\frac{X_1-\mathbf{x}}{h}), g(\frac{X_1-\mathbf{y}}{h})\} = o(h^p).$$

Combining these results, we obtain that

$$\text{Cov}\{G_n(\mathbf{x}), G_n(\mathbf{y})\} = o(1) + o(h^p).$$

The lemma is thus proved. □

Recall from (4.25),

$$\text{Var}\{G_n(\mathbf{x})\} = f(\mathbf{x})\int g(\mathbf{u})^2d\mathbf{u} + O(h^p).$$

As a consequence of Lemma 4.7, we have

$$\mathrm{Corr}\{G_n(\mathbf{x}), G_n(\mathbf{y})\} = o(1),$$

as $h \to 0, nh^p \to \infty$. That is, $G_n(\mathbf{x}), G_n(\mathbf{y})$ are asymptotically uncorrelated as $h \to 0, nh^p \to \infty$. We need a central limit theorem for martingale arrays for proof of Lemma 4.4.

CLT for martingale arrays. Now we state a martingale central limit theorem for martingale arrays which is used our proof of Lemma 4.4.

We state a CLT for square-integrable martingale differences (see Theorem 4, Chapter VII of Shiryayev (1984)). Given a double sequence on a probability space $(\Omega, \mathcal{F}, P)$:

$$(\xi_{ni}, \mathcal{F}_i^n), 0 \le i \le n, \;\; n \ge 1,$$

with $\xi_{n0} = 0, \mathcal{F}_0^n = \{\phi, \Omega\}, \mathcal{F}_i^n \subseteq \mathcal{F}_{i+1}^n \subseteq \mathcal{F}$, we require that ξ_{ni} is $\mathcal{F}_i^n$-measurable for each $n \ge 1, 1 \le i \le n$, i.e. $(\xi_{ni}, \mathcal{F}_i^n)$ is a stochastic sequence.

Proposition 4.1 *For each $n \ge 1$ let the stochastic sequence*

$$(\xi_{ni}, \mathcal{F}_i^n), 0 \le i \le n, \;\; n \ge 1,$$

be a square-integrable martingale difference:

$$E\xi_{ni}^2 < \infty, E(\xi_{ni}|\mathcal{F}_{i-1}^n) = 0.$$

Suppose that the Lindeberg condition is satisfied: there exists an $\epsilon > 0$ so that,

$$(L) \quad \sum_{i=1}^n E\{\xi_{ni}^2 1(|\xi_{ni}| > \epsilon)|\mathcal{F}_{i-1}^n\} \xrightarrow{\mathrm{P}} 0.$$

Then

$$\sum_{i=1}^n E\{\xi_{ni}^2|\mathcal{F}_{i-1}^n\} \xrightarrow{\mathrm{P}} \sigma^2 \Rightarrow \sum_{i=1}^n \xi_{ni} \xrightarrow{d} N(0, \sigma^2). \tag{4.27}$$

Now we prove the lemmas.

Proof of Lemmas 4.2, 4.3, 4.4.

PROOF OF LEMMA 4.2 AND LEMMA 4.3. Apply Lemma 4.6 to each element of

S_n, R_n, and going through same calculations as in the proof of Lemma 3.1 and Lemma 3.5, respectively. □

PROOF OF LEMMA 4.4.To establish asymptotic normality of vectors $\bar{Z}$, we need only to establish asymptotic normality for any linear combination of it by the Cramer-Wold device. Consider

$$l(\frac{X_i - \mathbf{x}}{h}) \triangleq \\ \{a + b^T(\frac{X_i - \mathbf{x}}{h}) + c^T \text{vech}\{(\frac{X_i - \mathbf{x}}{h})(\frac{X_i - \mathbf{x}}{h})^T\}\}K(\frac{X_i - \mathbf{x}}{h})\nu^{\frac{1}{2}}(X_i), \quad (4.28)$$

where a is any constant, b is any $p-$dimensional constant vector, c is any constant vector of dimension $p(p+1)/2$. Set

$$\xi_{ni} \triangleq \frac{1}{\sqrt{nh^p}} l(\frac{X_i - \mathbf{x}}{h})\varepsilon_i. \quad (4.29)$$

It is easy to check that $\{\xi_{ni}, \mathcal{F}_i^{XY}\}$ is a square-integrable martingale difference.

Now we check the Lindeberg condition:

$$\begin{aligned}
(L) \quad & \sum_{i=1}^{n} E\{\xi_{ni}^2 1(|\xi_{ni}| > \epsilon) \mid \mathcal{F}_{i-1}^{XY}\} \\
& \le \sum_{i=1}^{n} E\{\frac{|\xi_{ni}|^{2+\delta}}{\epsilon^\delta} \mid \mathcal{F}_{i-1}^{XY}\} \\
& = \sum_{i=1}^{n} \frac{1}{\epsilon^\delta} \frac{1}{(nh^p)^{\frac{2+\delta}{2}}} E\{|l(\frac{X_i - \mathbf{x}}{h})\varepsilon_i|^{2+\delta} \mid \mathcal{F}_{i-1}^{XY}\} \\
& = \frac{1}{(nh^p)^{1+\frac{\delta}{2}}\epsilon^\delta} \sum_{i=1}^{n} |l(\frac{X_i - \mathbf{x}}{h})|^{2+\delta} E\{|\varepsilon_i|^{2+\delta} \mid \mathcal{F}_{i-1}^{XY}\} \\
& = \frac{1}{(nh^p)^{1+\frac{\delta}{2}}\epsilon^\delta} E|\varepsilon_1|^{2+\delta} \sum_{i=1}^{n} |l(\frac{X_i - \mathbf{x}}{h})|^{2+\delta},
\end{aligned}$$

where the definition of $\mathcal{F}_{i-1}^{XY}$ and the assumptions that ε_i's are iid and are independent of $X_k, k \le i$ are used.

Applying Lemma 4.6 to $|l(\frac{X_i - \mathbf{x}}{h})|^{2+\delta}$, the above is equal to

$$\frac{1}{(nh^p)^{\frac{\delta}{2}}\epsilon^\delta} E|\varepsilon_1|^{2+\delta}\{f(\mathbf{x})\int |l(\mathbf{u})|^{2+\delta} d\mathbf{u} + O_p(h^p)\} \xrightarrow{\text{P}} 0,$$

if $h \to 0, nh^p \to \infty$, as $n \to \infty$. Here the finiteness of $\int |l(\mathbf{u})|^{2+\delta} d\mathbf{u}$ follows the fact that

$$l(\mathbf{u}) = (a + b^T\mathbf{u} + c^T \text{vech}\{\mathbf{u}\mathbf{u}^T\})K(\mathbf{u})$$

and thus $\int |l(\mathbf{u})|^4 d\mathbf{u} < \infty$ under the assumed moment condition on K. So the Lindeberg condition in Proposition 4.1 is satisfied if $h \to 0, nh^p \to \infty$.

Furthermore,

$$\sum_{i=1}^{n} E\{\xi_{ni}^2 \mid \mathcal{F}_{i-1}^{XY}\} = \frac{1}{nh^p} \sum_{i=1}^{n} l(\frac{X_i - \mathbf{x}}{h})^2.$$

Applying Lemma 4.6 to $|l(\frac{X_i - \mathbf{x}}{h})|^{2+\delta}$ again, the above is equal to

$$f(x) \int l(\mathbf{u})^2 d\mathbf{u} + O_p(h^p),$$

where

$$\int l(\mathbf{u})^2 d\mathbf{u} = \int \{a + b\mathbf{u} + c^T \text{vech}\{\mathbf{u}\mathbf{u}^T\}\}^2 K(\mathbf{u})^2 v(\mathbf{x} + h\mathbf{u}) d\mathbf{u}$$

$$= (a, b, c^T) v(\mathbf{x}) \int \begin{pmatrix} 1 \\ \mathbf{u} \\ \text{vech}\{\mathbf{u}\mathbf{u}^T\} \end{pmatrix} (1, \mathbf{u}^T, \text{vech}^T\{\mathbf{u}\mathbf{u}^T\}) K(u)^2 d\mathbf{u} (a, b, c^T)^T + O(h).$$

So by Proposition 4.1, we have

$$\sum_{i=1}^{n} \xi_{ni} \xrightarrow{d} N(0, (a, b, c^T) \Sigma_1 (a, b, c^T)^T),$$

where

$$\Sigma_1 = v(\mathbf{x}) f(\mathbf{x}) \int \begin{pmatrix} 1 \\ \mathbf{u} \\ \text{vech}\{\mathbf{u}\mathbf{u}^T\} \end{pmatrix} (1, \mathbf{u}^T, \text{vech}^T\{\mathbf{u}\mathbf{u}^T\}) K(\mathbf{u})^2 d\mathbf{u} + O(h).$$

By the Cramer-Wold device for proving joint asymptotic normality, this implies that

$$Z_n \xrightarrow{d} N(0, \Sigma_1).$$

So Lemma 4.4 is proved. □

Chapter 5

ESTIMATION OF LYAPUNOV EXPONENTS

5.1 Introduction

In this chapter, we consider estimations of global Lyapunov exponents (LE) and local Lyapunov exponents (LLE). The estimation of Lyapunov exponents and some challenging problems are discussed by McCaffrey et al (1992), Ellner et al (1991), Nychka et al (1992). The importance of LLE in identifying predictable regions in the phase space is discussed in Wolff (1992), and Bailey (1993). The partial derivative estimators are given by the locally weighted polynomial fit studied in Chapter 3 and Chapter 4. This chapter provides a systematic study of the estimators of LLE in a noisy system. The explicit results on the asymptotic bias and asymptotic covariance matrix of the LLE estimators may shed light on the estimation of LE.

We consider the following noisy model for a time series (without loss of generality we take $\tau = 1$):

$$x_{t+1} = m(x_t, \cdots, x_{t-p+1}) + \sigma\varepsilon_{t+1}, \tag{5.1}$$

where (A) $\varepsilon_1, \varepsilon_2, \ldots$ are iid random noises with zero mean and unit variance, and ε_{t+1}

is independent of $x_t, x_{t-1}, \ldots$.

Define the state space vector by the time delay method

$$X_t = (x_t, x_{t-1}, \cdots, x_{t-p+1})^T,$$

then we can write (5.1) in its state space form.

$$X_{t+1} = M(X_t) + GE_t, \tag{5.2}$$

where $M(\mathbf{x}) = (m(\mathbf{x}), x_p, \cdots, x_2)^T$ at a phase point $\mathbf{x} = (x_p, x_{p-1}, \cdots, x_1)^T$, and $G = \mathrm{diag}\{\sigma, 0, \cdots, 0\}, E_t = (\varepsilon_t, 0, \cdots, 0)^T$.

Some notations are needed later on. If $A = (a_{ij})$ is a $p \times p$ square matrix, $\mathrm{dg}\{A\}$ is the diagonal matrix from A by setting all supra- and infradiagonal elements in A equal to zero, that is

$$\mathrm{dg}\{A\} = \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{pp} \end{pmatrix}.$$

We also use $\mathrm{dgv}\{A\}$ to denote the vector consisting of the diagonal elements of A, i.e.

$$\mathrm{dgv}\{A\} = (a_{11} \cdots a_{pp})^T.$$

We denote a $p \times q$ zero matrix by $0_{p,q}$.

This chapter is organized as follows. The partial derivative estimation is applied to estimate the finite-step Jacobian product in Section 5.2. Then in Section 5.3, we will study the asymptotic theory of the eigenvalues from a random matrix, which is used to derive the corresponding theory of singular values in Section 5.4. The asymptotic results for the singular values of the estimated multistep Jacobian matrix is given in Section 5.5. In Section 5.6, the asymptotic theory of the local Lyapunov spectrum is given, and a method of constructing pointwise confidence intervals is prescribed.

5.2 Estimation of Jacobian Matrix

1. Nonparametric partial derivative estimation.

For the estimation of partial derivatives of m, besides assumption (A), we further make the following assumptions.

(B) Ergodicity. There is an invariant measure ρ for the Markov chains $\{X_i\}$. Furthermore, ρ is assumed ergodic and X_0 is sampled from ρ.

(C) Short-range dependence. Under (B), and we further assume that ρ is absolutely continuous and ε_1 has a density. As a result, the joint density of X_p, X_{p+j} exists for any $j \geq 1$, which is denoted by $f_j(\cdot,\cdot)$, and the marginal density is denoted by $f(\cdot)$. Note that the 'β_n' index is given by

$$\beta_n \triangleq \sup_{\mathbf{u},\mathbf{v}\in\mathcal{R}^p} \sum_{j=1}^{n} |f_j(\mathbf{u},\mathbf{v}) - f(\mathbf{u})f(\mathbf{v})|.$$

We assume $\beta_n = O(1)$.

(D) Moment Condition. There exists a $\delta > 0$ so that $E|\varepsilon|^{2+\delta}$ is finite.

(E) Smoothness. For an open neighborhood of $\mathbf{x}$, assume $m \in C^3(U), f \in C^0(U)$.

The local quadratic fit is used for estimating the partial derivatives. By Theorem 4.2 in Chapter 4, the local quadratic estimator has the following asymptotic normality property under conditions (A)-(E).

Corollary 5.1 *For l distinct interior points $\mathbf{x}_1, \ldots, \mathbf{x}_l$ inside the support of density f and $f(\mathbf{x}_j) > 0$ for all j, if there exist open neighborhoods U_i of $\mathbf{x}_i$ such that $m \in C^3(U_j), f \in C^0(U_j)$. Then the local quadratic partial derivative estimators $\widehat{D_m(\mathbf{x}_1)}, \ldots, \widehat{D_m(\mathbf{x}_l)}$ at each point $\mathbf{x}_1, \ldots, \mathbf{x}_l$ are asymptotically independent and normally distributed. Specifically, at each point, say $\mathbf{x} = (\mathbf{x_p}, \cdots, \mathbf{x_1})^{\mathbf{T}}$, we have for $h \to 0, nh^{p+2} \to \infty$ as $n \to \infty$,*

$$Z_n = (nh^{p+2})^{\frac{1}{2}}\{\widehat{D_m(\mathbf{x})} - D_m(\mathbf{x}) - b(\mathbf{x}, h)\} \xrightarrow{d} N(0, \Sigma), \tag{5.3}$$

where

$$b(\mathbf{x},\mathbf{h}) = \frac{h^2}{3!\mu_2}[b(\mathbf{x}) + o(1)], \tag{5.4}$$

and

$$\begin{aligned} b(\mathbf{x}) &= \int_{\mathcal{R}^p} u(\sum_{i=1}^{p} \frac{\partial}{\partial x_i} u_i)^3 m(\mathbf{x}) K(u) du \\ &= \begin{pmatrix} \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_p^3} + 3\mu_2^2 \sum_{i=1}^{p-1} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \partial x_p} \\ \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_{p-1}^3} + 3\mu_2^2 \sum_{i \neq p-1} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \partial x_{p-1}} \\ \vdots \\ \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_1^3} + 3\mu_2^2 \sum_{i=2}^{p} \frac{\partial^3 m(\mathbf{x})}{\partial x_i^2 \partial x_1} \end{pmatrix}, \end{aligned} \tag{5.5}$$

where

$$\Sigma = \frac{\sigma^2 J_2}{{\mu_2}^2 f(\mathbf{x})} I, \tag{5.6}$$

where $\mu_\ell = \int u_1^\ell K(\mathbf{u}) d\mathbf{u}, J_\ell = \int u_1^\ell K^2(\mathbf{u}) d\mathbf{u}$ *for nonnegative integers* ℓ*.*

2. Jacobian matrix. Some notations will be introduced and the partial derivative estimator is used to estimate the Jacobian matrix. Note that the Jacobian matrix of $M(\mathbf{x})$ is given by

$$T(\mathbf{x}) = \begin{pmatrix} & D_m^T(\mathbf{x}) & \\ I_{p-1} & & 0_{p-1,1} \end{pmatrix}, \tag{5.7}$$

where $D_m(\mathbf{x})$ is the partial derivative vector. An estimator of the Jacobian matrix is given by substituting $D_m(\mathbf{x})$ by $\hat{D}_m(\mathbf{x})$.

Noticing,

$$\hat{T}(\mathbf{x}) - T(\mathbf{x}) = \begin{pmatrix} \hat{D}_m^T(\mathbf{x}) - D_m^T(\mathbf{x}) \\ 0_{p-1,p} \end{pmatrix},$$

let

$$B(\mathbf{x}, h) = \begin{pmatrix} b(\mathbf{x}, h)^T \\ 0_{p-1,p} \end{pmatrix}, \tag{5.8}$$

where $b(\mathbf{x}, h)$ is given in (5.4), and

$$W(\mathbf{x}) = \frac{\sigma\sqrt{J_2}}{\mu_2\sqrt{f(\mathbf{x})}} \begin{pmatrix} Z^T \\ 0_{p-1,p} \end{pmatrix}, \tag{5.9}$$

where $Z \sim N(0, I)$. Then, by (5.3), we have as $h \rightarrow 0, nh^p \rightarrow \infty$,

$$(nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}) - T(\mathbf{x}) - B(\mathbf{x}, h)\} \xrightarrow{d} W(\mathbf{x}). \qquad (5.10)$$

Given l fixed and distinct points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ in embedded state space R^p, an estimator of the $l-$step Jacobian matrix product by $T^l = T(\mathbf{x}_l)\cdots T(\mathbf{x}_2)T(\mathbf{x}_1)$ is given by substituting $D_m(\mathbf{x}_i)$'s by $\hat{D}_m(\mathbf{x}_i)$'s. Our purpose in this section is to express the asymptotic properties of $\hat{T}^l$ for fixed l.

3.($l = 2$). Note that we have decomposition for $\hat{T}^2$

$$\begin{aligned}
&(nh^{p+2})^{1/2}[\hat{T}(\mathbf{x}_2)\hat{T}(\mathbf{x}_1) - \{T(\mathbf{x}_2) + B(\mathbf{x}_2, h)\}\{T(\mathbf{x}_1) + B(\mathbf{x}_1, h)\}] \\
&\quad = \; (nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}_2) - T(\mathbf{x}_2) - B(\mathbf{x}_2, h)\}T(\mathbf{x}_1) \\
&\qquad + T(\mathbf{x}_2)(nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}_1) - T(\mathbf{x}_1) - B(\mathbf{x}_1, h)\} \\
&\qquad + (nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}_2) - T(\mathbf{x}_2) - B(\mathbf{x}_2, h)\}B(\mathbf{x}_1, h) \\
&\qquad + B(\mathbf{x}_2, h)(nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}_1) - T(X_1) - B(\mathbf{x}_1, h)\} \\
&\qquad + (nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}_2) - T(\mathbf{x}_2) - B(\mathbf{x}_2, h)\}\{\hat{T}(\mathbf{x}_1) - T(\mathbf{x}_1) - B(\mathbf{x}_1, h)\},
\end{aligned}$$

Define the matrices $W(\mathbf{x}_1), W(\mathbf{x}_2)$ by

$$W_i = \frac{\sigma\sqrt{J_2}}{\mu_2\sqrt{f(\mathbf{x}_i)}} \begin{pmatrix} Z_i^T \\ 0_{p-1,p} \end{pmatrix}, \; i = 1, 2, \qquad (5.11)$$

where Z_1, Z_2 are iid $N(0, I)$.

As in (5.10), since

$$(nh^{p+2})^{1/2}\{\hat{T}(\mathbf{x}_i) - T(\mathbf{x}_i) - B(\mathbf{x}_i, h)\} \xrightarrow{d} W_i, i = 1, 2,$$

it follows from the preceding decomposition for $\hat{T}^2$ that as $h \rightarrow 0, nh^{p+2} \rightarrow \infty$,

$$\begin{aligned}
&(nh^{p+2})^{1/2}[\hat{T}^2 - T^2 - \{T(\mathbf{x}_2)B(\mathbf{x}_1, h) + B(\mathbf{x}_2, h)T(\mathbf{x}_1)\} + O(h^4)] \\
&\qquad\qquad \xrightarrow{d} W_2 T(\mathbf{x}_1) + T(\mathbf{x}_2)W_1. \qquad (5.12)
\end{aligned}$$

4. Any fixed l. Denote $\tilde{T}(\mathbf{x}_i) = T(\mathbf{x}_i) + B(\mathbf{x}_i, h)$, $1 \leq i \leq l$, where $b(\mathbf{x}_i, h)(i = 1, 2, \ldots, l)$ is defined as in (5.4), and $\tilde{T}^l = \tilde{T}(\mathbf{x}_l)\cdots\tilde{T}(\mathbf{x}_1)$.

Use the decomposition for matrix product

$$\begin{aligned}\hat{T}^l - \tilde{T}^l &= \hat{T}(\mathbf{x}_l)\cdots\hat{T}(\mathbf{x}_2)\hat{T}(\mathbf{x}_1) - \tilde{T}(\mathbf{x}_l)\cdots\tilde{T}(\mathbf{x}_2)\tilde{T}(\mathbf{x}_1)\\ &= \textstyle\sum_{j=1}^{l}\tilde{T}(\mathbf{x}_l)\cdots\tilde{T}(\mathbf{x}_{j+1})(\hat{T}(\mathbf{x}_j)-\tilde{T}(\mathbf{x}_j))\tilde{T}(\mathbf{x}_{j-1})\cdots\tilde{T}(\mathbf{x}_1)\\ &\quad + \textstyle\sum_{1\le k<j\le l}\tilde{T}(\mathbf{x}_l)\cdots\tilde{T}(\mathbf{x}_{j+1})(\hat{T}(\mathbf{x}_j)-\tilde{T}(\mathbf{x}_j))\\ &\quad \tilde{T}(\mathbf{x}_{j-1})\cdots\tilde{T}(\mathbf{x}_{k+1})(\hat{T}(\mathbf{x}_k)-\tilde{T}(\mathbf{x}_k))\tilde{T}(\mathbf{x}_{k-1})\cdots\tilde{T}(\mathbf{x}_1)\\ &\quad + \cdots\\ &\quad +(\hat{T}(\mathbf{x}_l)-\tilde{T}(\mathbf{x}_l))\cdots(\hat{T}(\mathbf{x}_2)-\tilde{T}(\mathbf{x}_2))(\hat{T}(\mathbf{x}_1)-\tilde{T}(\mathbf{x}_1)),\end{aligned}$$

where $T_{k+1}^{j} = T(\mathbf{x}_j)\cdots T(\mathbf{x}_{k+1}), k < j, T^{k-1} = T(\mathbf{x}_{k-1})\cdots T(\mathbf{x}_1)$, with the conventions that $T^0 = T_{l+1}^{l} = T(\mathbf{x}_0) = T(\mathbf{x}_{l+1}) = I$.

As in (5.10), since

$$(nh^{p+2})^{1/2}(\hat{T}(\mathbf{x}_i) - \tilde{T}(\mathbf{x}_i) \xrightarrow{d} W_i, i = 1, 2, \ldots, l$$

where $W_i, i = 1, 2, \ldots, l$ are independent, and defined similarly as in (5.11), we obtain as $h \to 0, nh^{p+2} \to \infty$,

$$\begin{aligned}(nh^{p+2})^{1/2}\{\hat{T}^l - \tilde{T}^l\} &= \sum_{j=1}^{l} T_{j+1}^{l}(nh^{p+2})^{1/2}(\hat{T}(\mathbf{x}_j) - \tilde{T}(\mathbf{x}_j))T^{j-1}\\ &\quad + O(h^2) + O_p(\{nh^{p+2}\}^{-1/2}),\end{aligned}$$

where $\tilde{T}^m$ has the approximation

$$\tilde{T}^l = T^l + \sum_{j=1}^{l} T_{j+1}^{l} B(\mathbf{x}_j, h)T^{j-1} + O(h^4).$$

In all, we have the following corollary.

Corollary 5.2 *Under conditions of Corollary 5.1, for fixed l, we have as $h \to 0, nh^{p+2} \to \infty$,*

$$(nh^{p+2})^{1/2}\{\hat{T}^l - T^l - \sum_{j=1}^{l} T_{j+1}^{l} B(\mathbf{x}_j, h)T^{j-1} - O(h^4)\} \xrightarrow{d} \sum_{j=1}^{l} T_{j+1}^{l} W_j T^{j-1}. \quad (5.13)$$

In order to derive corresponding results for the estimators of local Lyapunov exponents from those of the Jacobian products given above, we use the asymptotic theory of the eigenvalues and the singular values from a random matrix, which is studied in next two sections.

5.3 Asymptotic Distribution of Eigenvalues

The delta method. In various contexts in multivariate analysis, the asymptotic distribution of the eigenvalues of a random matrix is needed. The problem is simple under the assumption that the eigenvalues of the limiting matrix have multiplicity one, since a Taylor expansion of the characteristic equation at the eigenvalues can be given. The method using the Taylor expansion is usually called the *delta method* in the literature. In this section, we will give an exposition of the delta method in full generality, expanding on a derivation given by Richard Smith, who assumes that all the eigenvalues of the limiting matrix have multiplicity one.

Expansion for the eigenvalues having multiplicity more than one is more complicated. In the case of symmetric matrices, an alternative approach introduced by Eaton and Tyler (1991) using the Wielandt's eigenvalue inequality can be used to circumvent the complications in the case of multiple roots.

For an arbitrary matrix X, its eigenvalues λ as an implicit function of X are defined as solutions to the characteristic equation

$$f(X,\lambda) = |X - \lambda I| = 0. \tag{5.14}$$

The multiplicity of a root of the characteristic equation is called the multiplicity of the eigenvalue. Denote the eigenvalues of X by $\varphi(X) = (\varphi_1(X), \ldots, \varphi_p(X))^T$ (some are possibly complex numbers).

One application is the following: suppose that X_n is a $p \times p$ random matrix satisfying

$$c_n^{1/2}(X_n - A) \xrightarrow{d} W, \tag{5.15}$$

where A is nonrandom, W is a random matrix, for an increasing sequence $c_n \to \infty$. We are interested in finding the asymptotic distribution of

$$Y_{n,k} = c_n^{1/2}\{\varphi_k(X_n) - d_k\}, \text{ where } d_k = \varphi_k(A), 1 \le k \le p, \tag{5.16}$$

or their joint asymptotic distribution

$$Y_n = (Y_{n,1}, \ldots, Y_{n,p})^T = c_n^{1/2}(\varphi(X_n) - \varphi(A)).$$

The simple root case. The idea is simple. If $\varphi_k(X)$ is a differentiable function of X, it then follows from the multivariate Taylor expansion that $\lambda = \varphi_k(X)$ can be expanded near $d_k = \varphi_k(A)$, for X being close to A. So the main step in using the delta method is to justify the differentiability property of $\varphi_k(X)$'s in a neighborhood of matrix A. The implicit function theorem in multivariate calculus will be used and a version is stated here.

Theorem 5.1 (Implicit Function Theorem) *Given a $(q+1)-$variate function $F(\mathbf{x}, y) = F(x_1, x_2, \ldots, x_q, y)$, where $\mathbf{x} = (x_1, \ldots, x_q)$, let $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,q})$, and U be a rectangle in R^{q+1} centered at a point $(\mathbf{x}_0, y_0) = (x_{0,1}, \ldots, x_{0,q}, y_0)$. Let $C^k(U)$ denote the class of functions defined on U which are $k-$times continuously differentiable. Assume that $F(\mathbf{x}, y)$ satisfies the following*

1. $F(\mathbf{x}, y) \in C^k(U)$,
2. $F(\mathbf{x}_0, y_0) = 0$,
3. $\frac{\partial F(\mathbf{x}_0, y_0)}{\partial y} \neq 0$.

Then, there exist $\eta > 0$, and $\delta = (\delta_1, \ldots, \delta_q)$, each $\delta_i > 0$, so that for any $\mathbf{x} \in (\mathbf{x}_0 - \delta, \mathbf{x}_0 + \delta) \subset R^q$, there is a uniquely defined function $y = \varphi(\mathbf{x}) = \varphi(x_1, \ldots, x_q)$ which satisfies

(A) $F(\mathbf{x}, \varphi(\mathbf{x})) = 0$, *for all* $\mathbf{x} \in (\mathbf{x}_0 - \delta, \mathbf{x}_0 + \delta)$,

(B) $|\varphi(\mathbf{x}) - y_0| < \eta$,

(C) $\varphi(\mathbf{x}_0) = y_0$,

(D) $\varphi(\mathbf{x}) \in C^k(\mathbf{x}_0 - \delta, \mathbf{x}_0 + \delta)$, *and furthermore*

$$\frac{\partial \varphi(\mathbf{x})}{\partial x_i} = - \left. \frac{\partial F(\mathbf{x}, y)}{\partial x_i} \middle/ \frac{F(\mathbf{x}, y)}{\partial y} \right|_{y=\varphi(\mathbf{x})}, \quad i = 1, 2, \ldots, q.$$

The following general lemma is obtained as an immediate application of the implicit function theorem (Theorem 5.1).

Lemma 5.1 *Let $\varphi_k(X)$ denote the kth eigenvalue of a $p \times p$ matrix X. For any given matrix A with the kth eigenvalue $d_k = \varphi_k(A)$ having multiplicity one, then $\varphi_k(X)$ is a uniquely well-defined function in a neighborhood of A, and is differentiable up to any order. Moreover,*

$$\left.\frac{\partial \varphi_k(X)}{\partial x_{ij}}\right|_{X=A} = -\frac{f_{ij}(A, d_k)}{f'(A, d_k)},$$

where

$$f_{ij}(A, d_k) = \frac{\partial f(X, \lambda)}{\partial x_{ij}}|_{X=A,\lambda=d_k},$$

$$f'(A, d_k) = -\prod_{i \neq k}(d_i - d_k)^{p_i}.$$

Proof: Apply the implicit function theorem (Theorem 5.1) to $f(X, \lambda)$ given in (5.14), with initial values given by $\mathbf{x}_0 =$ elements of $A, y_0 = d_k$. Since $f(X, \lambda)$ is a smooth function of X and λ, the differentiability of the eigenvalue, say $\varphi_k(X)$, in a neighborhood of A depends on the fulfillment of condition (3) in the implicit function theorem

$$\left.\frac{\partial f(X, \lambda)}{\partial \lambda}\right|_{X=A,\lambda=d_k} \neq 0. \tag{5.17}$$

Denote

$$f_{ij}(X, \lambda) = \frac{\partial f(X, \lambda)}{\partial x_{ij}}, f'(X, \lambda) = \frac{\partial f(X, \lambda)}{\partial \lambda}.$$

Recall that, any matrix A can be reduced to its Jordan form Λ in $\mathcal{C}$ by a similar transformation, that is, there exists a nonsingular matrix V such that

$$V^{-1}AV = \Lambda,$$

where

$$\Lambda = \text{diag}\{\Lambda_1, \ldots, \Lambda_r\}, \tag{5.18}$$

where Λ_s has the form

$$\Lambda_s = \begin{pmatrix} d_s & 1 & 0 & \cdots & 0 \\ 0 & d_s & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_s \end{pmatrix}, \tag{5.19}$$

where d_s is an eigenvalue of A. Denote the $l(l \leq r)$ distinct eigenvalues of A by $d_1, \ldots, d_l$, with each d_s of multiplicity $p_s, 1 \leq s \leq l, p_1 + \ldots + p_l = p$

The characteristic equation of A is given by

$$f(A, \lambda) = f(\Lambda, \lambda) = \prod_{s=1}^{l} (d_s - \lambda)^{p_s}. \tag{5.20}$$

Since each d_s is independent of λ,

$$f'(A, \lambda) = -\sum_{s=1}^{l} p_s (d_s - \lambda)^{p_s - 1} \prod_{i \neq s} (d_i - \lambda)^{p_i}.$$

Thus,

$$f'(A, d_k) = \begin{cases} -\prod_{i \neq k} (d_i - d_k)^{p_i} & \text{if } p_k = 1 \\ 0 & \text{if } p_k > 1 \end{cases}. \tag{5.21}$$

So $f'(A, d_k) \neq 0$ if and only if d_k has multiplicity one, that is, $p_k = 1$. Since the conditions of the implicit function theorem are satisfied, it follows that the eigenvalue $\varphi_k(X)$ is differentiable up to any order , that is, it is analytic, in a neighborhood of the matrix A if d_k has multiplicity one. The lemma is proved. □

We need the following lemma.

Lemma 5.2 *Given that X has entries $(x_{ij}), 1 \leq i, j \leq p$, let X^{ij} denote the cofactor of x_{ij}. Then,*

$$\frac{\partial |X|}{\partial x_{ij}} = X^{ij}.$$

PROOF. The expansion of $|X|$ by elements of the ith row is

$$|X| = \sum_{h=1}^{p} x_{ih} X^{ih}.$$

Since X_{ih} does not contain x_{ij}, the lemma follows. □

In summary, we have the following corollary.

Corollary 5.3 *Assume (5.15) and that the kth eigenvalue $\varphi_k(A)$ has multiplicity one, then*

$$c_n^{1/2} (\varphi_k(X_n) - d_k) \xrightarrow{d} \sum_{i,j} \varphi_k(i, j) w_{ij},$$

where $\varphi_k(i, j)$ is given in (5.23).

Proof: Applying Lemma 5.2 to evaluate $f_{ij}(A, d_k)$, denoting the cofactor of the (i, j)th element in matrix $A - d_k I$ as $(A - d_k I)^{ij}$, we have

$$f_{ij}(A, d_k) = (A - d_k I)^{ij}. \tag{5.22}$$

Lemma 5.1 implies that, if d_k has multiplicity one, by using (5.21),

$$\left.\frac{\partial \varphi_k(X)}{\partial x_{ij}}\right|_{X=A} = \frac{(A - d_k I)^{ij}}{\prod_{i \neq k}(d_i - d_k)^{p_i}} \triangleq \varphi_k(i, j). \tag{5.23}$$

Note that if $d_k = \varphi_k(A)$ has multiplicity one, denoting $X = (x_{ij}), A = (a_{ij})$, and $\|\cdot\|$ be a matrix norm, the Taylor expansion for $\varphi_k(X)$ at x being near A is given by

$$\varphi_k(X) - d_k = \sum_{i,j} \varphi_k(i, j)(x_{ij} - a_{ij}) + o(\|X - A\|),$$

where $\varphi_k(i, j)$ is given in (5.23). From (5.15), denoting $X_n = (x_{n,ij}), W = (w_{ij})$, it follows that

$$c_n^{1/2}(\varphi_k(X_n) - d_k) = \sum_{i,j} \varphi_k(i, j) c_n^{1/2}(x_{n,ij} - a_{ij}) + o_p(1),$$

and

$$c_n^{1/2}(\varphi_k(X_n) - d_k) \xrightarrow{d} \sum_{i,j} \varphi_k(i, j) w_{ij}.$$

End of proof. □

Now let's see how to simplify calculations in (5.23) or (5.22). Since any similarity transformation to assumption (5.15) will not change (5.16), without loss of generality we can assume that A is of the Jordan form Λ, which will simplify (5.23) considerably. For simplicity, in the rest of this section we will consider the case that A is a diagonal matrix, that is,

$$A = \mathrm{diag}\{d_1, \ldots, d_p\}. \tag{5.24}$$

where $d_k = \varphi_k(A)$. Note that A is diagonalizable under a similarity transformation, if all the eigenvalues of A have multiplicity one, or when A is a symmetric matrix. We have the following theorem.

Theorem 5.2 *Given $p \times p$ matrices X_n, A which satisfy (5.15), and assume that A is diagonalizable, that is there exists a nonsingular matrix V such that*

$$V^{-1}AV = \mathrm{diag}\{d_1, \cdots, d_p\}.$$

Denote

$$U = V^{-1} = \begin{pmatrix} U_1^T \\ \vdots \\ U_p^T \end{pmatrix}, V = (V_1, \cdots, V_p),$$

let $\varphi_k(X_n)$ denote the kth eigenvalue of matrix X_n, then if d_k has multiplicity one, the asymptotic distribution of $\varphi_k(X_n)$ is given by

$$Y_{n,k} = c_n^{1/2}\{\varphi_k(X_n) - d_k\} \xrightarrow{d} U_k^T W V_k. \tag{5.25}$$

If all the d_k's have multiplicity one, the joint asymptotic distribution of the vector $\varphi(X_n) = (\varphi_1(X_n), \ldots, \varphi_p(X_n))^T$ is given by

$$Y_n = c_n^{1/2}\{\varphi(X_n) - \varphi(A)\} \xrightarrow{d} (U_1^T W V_1, \ldots, U_p^T W V_p)^T.$$

Proof: Evaluating (5.22) in the case $A = \mathrm{diag}\{d_1, \ldots, d_p\}$, we have

$$f_{ij}(A, \lambda) = \begin{cases} \prod_{l \neq i}(d_l - \lambda) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Thus,

$$\begin{aligned} f_{ij}(A, d_k) &= \begin{cases} \prod_{l \neq i}(d_l - d_k) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \\ &= \begin{cases} \prod_{l \neq k}(d_l - d_k) & \text{if } i = j = k \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

From (5.21),

$$f'(A, d_k) = -\prod_{l \neq k}(d_l - d_k).$$

By Lemma 5.1,

$$\left.\frac{\partial \varphi_k(X)}{\partial x_{ij}}\right|_{X=A} = \begin{cases} 1 & \text{if i=j=k} \\ 0 & \text{otherwise} \end{cases},$$

that is,

$$\left.\frac{\partial \varphi_k(X)}{\partial X}\right|_{X=A} = E_{kk}, \tag{5.26}$$

where E_{kk} is the matrix whose kth diagonal element is one and all other elements are zeros.

Using Taylor expansion for $\varphi_k(X)$ at X near A and (5.26), we have obtained the theorem. □

A more general setting. Assume that an asymmetric matrix X_n satisfies

$$c_n^{1/2}(X_n - A_n) \xrightarrow{d} W, \text{ where } W \text{ is a random matrix,} \tag{5.27}$$

where c_n is an increasing sequence and $c_n \to \infty, A_n$ is a convergent sequence of matrices such that

$$A_n = A + B_n,\ A \text{ is nonrandom },B_n \xrightarrow{p} 0. \tag{5.28}$$

We have obtained the following corollary.

Corollary 5.4 *Assume the general setup (5.27), under the assumptions on A and notations of Theorem 5.2,*

$$Y_{n,k} = c_n^{1/2}\{\varphi_k(X_n) - d_k - U_k^T B_n V_k - o(B_n)\} \xrightarrow{d} U_k^T W V_k.$$

If all the d_k's have multiplicity one, the joint asymptotic distribution is given by

$$Y_n = c_n^{1/2}\{\varphi(X_n) - \varphi(A) - (U_1^T B_n V_1, \cdots, U_p^T B_n V_p)^T - o(B_n)\}$$

$$\xrightarrow{d} (U_1^T W V_1, \ldots, U_p^T W V_p)^T.$$

Proof: Under the assumption on A in Theorem 5.2, and using the same notations, we have that using the Taylor expansion,

$$c_n^{1/2}\{\varphi_k(X_n) - \varphi_k(A_n)\} = \sum_{i,j} \frac{\partial \varphi_k(C_n)}{\partial x_{ij}} U_i^T \{c_n^{1/2}(X_n - A_n)\} V_j + o_p(\|X_n - A_n\|), \tag{5.29}$$

where C_n is a matrix between $U^T X_n V$ and $U^T A_n V$, that is,

$$\|C_n - U^T A_n V\| \le \|U^T (X_n - A_n) V\|.$$

From (5.27) and (5.28), it follows that

$$C_n \xrightarrow{p} U^T A V = \text{diag}\{d_1, \cdots, d_p\} \triangleq D.$$

So by the continuity of $\frac{\partial \varphi_k(X)}{\partial x_{ij}}$ in the neighborhood of D,

$$\frac{\partial \varphi_k(C_n)}{\partial x_{ij}} \xrightarrow{p} \frac{\partial \varphi_k(D)}{\partial x_{ij}} = E_{kk}.$$

Thus, it follows from (5.29) that,

$$c_n^{1/2}\{\varphi_k(X_n) - \varphi_k(A_n)\} \xrightarrow{d} U_k^T W V_k,$$

where,

$$\varphi_k(A_n) = d_k + U_k^T B_n V_k + o(B_n)$$

from (5.28). The rest of the lemma follows similarly. □

The case of multiple roots. The delta or perturbation method works well for the simple root case since the eigenvalues are analytic about a simple root. However, there seems to be a limitation for the delta method to be applied to multiple root case. The eigenvalues, although continuous, are not differentiable at points of multiple roots. Let's see an example.

Example. Consider the 2×2 matrix

$$X = \begin{pmatrix} x & \epsilon \\ \epsilon & x \end{pmatrix}.$$

It is easy to see that,

$$\varphi_1(X) = x + |\epsilon|, \varphi_2(X) = x - |\epsilon|,$$

which are continuous but not differentiable at $\epsilon = 0$. The reason lies in the fact that X is a perturbation of a diagonal matrix $A = xI$ which has a multiple root x.

As a result, the delta method can not be used directly, and the expansions of the eigenvalues at multiple roots are more complicated. In the case of symmetric matrices, Eaton and Tyler (1991) have employed the Wielandt's inequality to circumvent the complications caused by the multiplicity of roots by considering the submatrix corresponding to the multiplicity of the roots. We refer to Eaton and Tyler (1991) for more details for the multiple root case.

5.4 Asymptotic Theory of Singular Values

In this section, we consider the asymptotic theory of the singular values from a random matrix, as a direct application of the eigenvalue theory discussed in last section. We will use the general setup (5.27) and (5.28).

Denote the singular values of a matrix X by $\delta_1(X) \geq \ldots \geq \delta_p(X)$, which are defined by $\varphi(\{X^TX\}^{1/2})$. Our purpose is to study the asymptotic behavior of

$$Y_n = (\delta_1(X_n), \cdots, \delta_p(X_n))^T - (\delta_1(A), \cdots, \delta_p(A))^T. \tag{5.30}$$

From (5.27) and (5.28), we have

$$c_n^{1/2}(X_n^TX_n - A_n^TA_n) \xrightarrow{d} W^TA + A^TW \triangleq \bar{W}, \tag{5.31}$$

where $A_n^TA_n$ has the approximation

$$A_n^TA_n = A^TA + B_n^TA + A^TB_n + o(B_n).$$

We have the following theorem.

Theorem 5.3 *Assume that all the singular values of A have multiplicity one (otherwise consider only those singular values which have multiplicity one), then*

$$c_n^{1/2}\left\{\begin{pmatrix} \delta_1(X_n) \\ \vdots \\ \delta_p(X_n) \end{pmatrix} - \begin{pmatrix} \delta_1(A) \\ \vdots \\ \delta_p(A) \end{pmatrix} - \begin{pmatrix} U_1^TB_nV_1 \\ \vdots \\ U_p^TB_nV_p \end{pmatrix} - o(B_n)\right\} \xrightarrow{d} \begin{pmatrix} U_1^TWV_1 \\ \vdots \\ U_p^TWV_p \end{pmatrix} \tag{5.32}$$

Proof: There exists a singular value decomposition (SVD) of A, i.e. there exist orthogonal matrices U, V such that

$$A = U\Delta V^T, \text{ where } \Delta = \mathrm{diag}\{\delta_1(A), \ldots, \delta_p(A)\}. \tag{5.33}$$

Consequently, $AV = U\Delta$, and $V^T A^T AV = \mathrm{diag}\{\delta_1^2(A), \ldots, \delta_p^2(A)\}$.

For simplicity, we will assume that all the singular values of A have multiplicity one (or consider only the singular values which have multiplicity one), then apply Corollary 5.4 to (5.31), we obtain

$$\begin{aligned} & c_n^{1/2}\{\varphi(X_n^T X_n) - \varphi(A_n^T A_n)\} \\ & \quad \xrightarrow{d} (V_1^T \bar{W} V_1, \cdots, V_p^T \bar{W} V_p)^T = \mathrm{dgv}\{V^T \bar{W} V\} \end{aligned} \tag{5.34}$$

where dgv denotes the vector consisting of the diagonal elements (see Section 5.1), $V = (V_1, \cdots, V_p)$, and $\varphi(A_n^T A_n)$ has the approximation

$$\varphi(A_n^T A_n) = \varphi(A^T A) + \mathrm{dgv}\{V^T(B_n^T A + A^T B_n)V\}. \tag{5.35}$$

Furthermore, using $AV = U\Delta$ and noting that Δ is a diagonal matrix,

$$\begin{aligned} \mathrm{dgv}\{V^T \bar{W} V\} &= \mathrm{dgv}\{V^T W^T AV + V^T A^T W V\} \\ &= \mathrm{dgv}\{V^T W^T U\Delta + \Delta U^T W V\} \\ &= 2\Delta \mathrm{dgv}(U^T W V). \end{aligned}$$

Similarly,

$$\mathrm{dgv}\{V^T(B_n^T A + A^T B_n)V\} = 2\Delta \mathrm{dgv}(U^T B_n V).$$

Denote $U = (U_1, \cdots, U_p)$. By using transformation $y = \sqrt{x}$, the theorem is then proved from (5.34) and (5.35). □

5.5 Application to Estimation of Singular Values of Jacobian Product.

Note that $T^l, \hat{T}^l$ are defined and studied in Section 5.2. Denote

$$\hat{\delta}_i(l) = \delta_i(\hat{T}^l), \delta_i(l) = \delta_i(T^l), 1 \le i \le p,$$

and denote the orthogonal matrices $U(l), V(l)$ in the SVD

$$T^l = U(l)\mathrm{diag}\{\delta_1(l), \cdots, \delta(l)\}V^T(l).$$

For simplicity we assume that the singular values of T^l have multiplicity one. Otherwise we will consider only those singular values which have multiplicity one. We have the following theorem.

Theorem 5.4 *Assume that all the singular values of T^l have multiplicity one, then under conditions Corollary 5.1, we have*

$$(nh^{p+2})^{1/2}\left\{\begin{pmatrix} \hat{\delta}_1(l) \\ \vdots \\ \hat{\delta}_p(l) \end{pmatrix} - \begin{pmatrix} \delta_1(l) \\ \vdots \\ \delta_p(l) \end{pmatrix}\right.$$
$$\left. -\frac{h^2}{6\mu_2}\sum_{j=1}^{l}\begin{pmatrix} \{U_1^T(l)T_{j+1}^l(1)\}\{b^T(\mathbf{x}_j)T^{j-1}V_1(l)\} \\ \vdots \\ \{U_p^T(l)T_{j+1}^l(1)\}\{b^T(\mathbf{x}_j)T^{j-1}V_p(l)\} \end{pmatrix} - o(h^2)\right\} \xrightarrow{d} N(0, \Sigma^s),$$

where $\Sigma^s = (\sigma_{ij}^s)$,

$$\sigma_{ii}^s = \frac{\sigma^2 J_2}{\mu_2^2}\sum_{k=1}^{l}\frac{1}{f(\mathbf{x}_k)}\{U_i^T(l)T_{k+1}^l(1)\}^2\{V_i^T(l)(T^{k-1})^T T^{k-1}V_i(l)\},$$

$$\sigma_{ij}^s = \frac{\sigma^2 J_2}{\mu_2^2}\sum_{k=1}^{l}\frac{1}{f(\mathbf{x}_k)}\{U_i^T(l)T_{k+1}^l(1)\}\{U_j^T(l)T_{k+1}^l(1)\}\{V_i^T(l)(T^{k-1})^T T^{k-1}V_j(l)\},$$

for $1 \le i \le p, 1 \le j \le p$.

Proof: From Corollary 5.1 and Theorem 5.3, we have

$$
\begin{aligned}
&(nh^{p+2})^{1/2}\left\{\begin{pmatrix}\hat{\delta}_1(l)\\ \vdots\\ \hat{\delta}_p(l)\end{pmatrix}-\begin{pmatrix}\delta_1(l)\\ \vdots\\ \delta_p(l)\end{pmatrix}\right.\\
&\left.-\begin{pmatrix}\sum_{j=1}^{l}U_1^T(l)T_{j+1}^{l}B(\mathbf{x}_j,h)T^{j-1}V_1(l)\\ \vdots\\ \sum_{j=1}^{l}U_p^T(l)T_{j+1}^{l}B(\mathbf{x}_j,h)T^{j-1}V_p(l)\end{pmatrix}-o(h^2)\right\}\\
&\xrightarrow{d}\begin{pmatrix}\sum_{j=1}^{l}U_1^T(l)T_{j+1}^{l}W_jT^{j-1}V_1(l)\\ \vdots\\ \sum_{j=1}^{l}U_p^T(l)T_{j+1}^{l}W_jT^{j-1}V_p(l)\end{pmatrix},
\end{aligned}
$$

where

$$U(l)=(U_1(l),\cdots,U_p(l)),V(l)=(V_1(l),\cdots,V_p(l)).$$

Furthermore, denote

$$T_{j+1}^{l}=(T_{j+1}^{l}(1),\cdots,T_{j+1}^{l}(p)). \tag{5.36}$$

From the definition of $B(\mathbf{x},h)$ in (5.8), (5.4), it is seen that

$$T_{j+1}^{l}B(\mathbf{x}_j,h)=T_{j+1}^{l}(1)b^T(\mathbf{x}_j,h)=\frac{h^2}{3!\mu_2}T_{j+1}^{l}(1)b(\mathbf{x}_j)+o(h^2),$$

and

$$
\begin{aligned}
&U_i^T(l)T_{j+1}^{l}B(\mathbf{x}_j,h)T^{j-1}V_i(l)=\\
&\quad\frac{h^2}{3!\mu_2}\{U_i^T(l)T_{j+1}^{l}(1)\}\{b^T(\mathbf{x}_j)T^{j-1}V_i(l)\},\ \text{for } 1\le i\le p.
\end{aligned}
$$

Similarly, from the definition of W_j in (5.11),

$$T_{j+1}^{l}W_j=\frac{\sigma\sqrt{J_2}}{\mu_2\sqrt{f(\mathbf{x}_j)}}T_{j+1}^{l}(1)Z_j^T,$$

and

$$U_i^T(l)T_{j+1}^{l}W_jT^{j-1}V_i(l)=\frac{\sigma\sqrt{J_2}}{\mu_2\sqrt{f(\mathbf{x}_j)}}\{U_i^T(l)T_{j+1}^{l}(1)\}\{Z_j^TT^{j-1}V_i(l)\},$$

for $1\le i\le p, 1\le j\le l$.

So we have proved

$$\begin{aligned}&(nh^{p+2})^{1/2}\left\{\begin{pmatrix}\hat{\delta}_1(l)\\ \vdots\\ \hat{\delta}_p(l)\end{pmatrix}-\begin{pmatrix}\delta_1(l)\\ \vdots\\ \delta_p(l)\end{pmatrix}\right.\\ &\left.-\frac{h^2}{6\mu_2}\sum_{j=1}^{l}\begin{pmatrix}\{U_1^T(l)T_{j+1}^l(1)\}\{b^T(\mathbf{x}_j)T^{j-1}V_1(l)\}\\ \vdots\\ \{U_p^T(l)T_{j+1}^l(1)\}\{b^T(\mathbf{x}_j)T^{j-1}V_p(l)\}\end{pmatrix}-o(h^2)\right\}\\ &\xrightarrow{d}\frac{\sigma\sqrt{J_2}}{\mu_2}\sum_{j=1}^{l}\frac{1}{\sqrt{f(\mathbf{x}_j)}}\begin{pmatrix}\{U_1^T(l)T_{j+1}^l(1)\}\{Z_j^T T^{j-1}V_1(l)\}\\ \vdots\\ \{U_p^T(l)T_{j+1}^l(1)\}\{Z_j^T T^{j-1}V_p(l)\}\end{pmatrix},\end{aligned}\tag{5.37}$$

which gives a characterization of the asymptotic distribution., from which the theorem follows easily. □

5.6 Estimation of Local Lyapunov Exponents

1. Asymptotic distribution of the LLE estimators. In this section, the results of Section 5.5 will be applied to estimation of the local Lyapunov spectrum. Note that the vector of the $l-$step local Lyapunov spectrum at a point $\mathbf{x}$ is defined by

$$(\lambda_1(l),\cdots,\lambda_p(l))^T \triangleq \frac{1}{l}(\log\{\delta_1(l)\},\ldots,\log\{\delta_p(l)\})^T,$$

and an estimator is given by

$$(\hat{\lambda}_1(l),\cdots,\hat{\lambda}_p(l))^T \triangleq \frac{1}{l}(\log\{\hat{\delta}_1(l)\},\ldots,\log\{\hat{\delta}_p(l)\})^T.$$

Apply Theorem 5.4 and the "delta method" for $\log x$, we have the following theorem.

Theorem 5.5 *Under conditions of Theorem 5.4, and if $\delta_p(l) \neq 0$,*

$$(nh^{p+2})^{1/2}\left\{\begin{pmatrix}\hat{\lambda}_1(l)\\ \vdots\\ \hat{\lambda}_p(l)\end{pmatrix}-\begin{pmatrix}\lambda_1(l)\\ \vdots\\ \lambda_p(l)\end{pmatrix}\right.$$

$$
-\frac{h^2}{6\mu_2 l}\begin{pmatrix} \delta_1^{-1}(l)\sum_{k=1}^{l}\{U_1^T(l)T_{k+1}^l(1)\}\{b^T(\mathbf{x}_k)T^{k-1}V_1(l)\} \\ \vdots \\ \delta_p^{-1}(l)\sum_{k=1}^{l}\{U_p^T(l)T_{k+1}^l(1)\}\{b^T(\mathbf{x}_k)T^{k-1}V_p(l)\} \end{pmatrix} - o(h^2)\Bigg\}
$$

$$
\xrightarrow{d} N(0, \Sigma^L),
$$

where $\Sigma^L = (\sigma_{ij}^L)$,

$$
\sigma_{ii}^L = \frac{\sigma^2 J_2}{l^2\mu_2^2\delta_i^2(l)}\sum_{k=1}^{l}\frac{1}{f(\mathbf{x}_k)}\{U_i^T(l)T_{k+1}^l(1)\}^2\|T^{k-1}V_i(l)\|^2,
$$

$$
\begin{aligned}
\sigma_{ij}^L &= \frac{\sigma^2 J_2}{l^2\mu_2^2\delta_i(l)\delta_j(l)} \\
&\quad \sum_{k=1}^{l}\frac{1}{f(\mathbf{x}_k)}\{U_i^T(l)T_{k+1}^l(1)\}\{U_j^T(l)T_{k+1}^l(1)\}\{V_i^T(l)(T^{k-1})^T T^{k-1}V_j(l)\},
\end{aligned}
$$

for $1 \le i \le p, 1 \le j \le p$.

2. Constructing Pointwise Confidence Intervals. Theorem 5.5 characterizes theoretically the variability associated with the local Lyapunov exponent estimators. In practice, Theorem 5.5 may be used to construct confidence intervals for the local Lyapunov exponents. One immediate difficulty in applying this theorem is that the estimators have asymptotic bias which may not be negligible. One way out of this is to choose the bandwidth h small enough to make the bias negligible, that is, to make $O(h^2)$ of smaller order than $(nh^{p+2})^{-1/2}$. This is the idea often used to construct confidence intervals and confidence region for a nonparametric regression in the literature. However, if the optimal bandwidth h is chosen so that it achieves the tradeoff between the asymptotic bias and the asymptotic variance, that is, when $O(h^2)$ and $(nh^{p+2})^{-1/2}$ are of the same order, there will always be asymptotic bias associated with the estimators. In order to give honest confidence intervals, we propose to use the "bias adjustment" method. Some pilot estimators for the third-order partial derivatives are required to give consistent estimates of the asymptotic bias, which are then used to adjust the bias of the local Lyapunov exponent estimators.

Specifically, for the estimation of the asymptotic bias, estimators for $b(\mathbf{x}_k), k = 1, 2, \ldots, l$ are needed, note that $b(\mathbf{x}_k)$'s are defined in (5.5) as functions of the third-order partial derivatives of the autoregression at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$. One simple method of deriving consistent estimators for the third-order partial derivatives of a regression function is the local cubic polynomial fit.

In addition, under the assumption that all singular values have multiplicity one, it is easy to see that $\hat{U}(l)$ and $\hat{V}(l)$, given by the singular value decomposition of $\hat{T}^l$, are consistent estimators of $U(l), V(l)$. A consistent estimator for the asymptotic bias of $\hat{\lambda}_i$ is given by $\frac{h^2}{6\mu_2 l}\hat{\beta}_i$, where

$$\hat{\beta}_i \triangleq \ \hat{\delta}_i^{-1}(l)\sum_{k=1}^{l}\{\hat{U}_k^T(l)\hat{T}_{k+1}^l(1)\}\{\hat{b}^T(\mathbf{x}_k)\hat{T}^{k-1}\hat{V}_k(l)\} \tag{5.38}$$

For the estimation of the asymptotic variance, consistent estimators for σ^2 and $f(\mathbf{x}_j), j = 1, 2, \ldots, l$ are needed. Natural consistent estimators for $f(\mathbf{x}_j)$'s are given by the kernel estimators, given by

$$\hat{f}(\mathbf{x}_j) = \frac{1}{nh_2^p}\sum_{i \neq j} K(\frac{X_i - \mathbf{x}_j}{h_2}), j = 1, 2, \ldots, l$$

where K is a kernel function.

Any consistent estimator $\hat{\sigma}$ will serve our purpose. There are several proposed estimators in the literature, of which the simplest one may be the differenced estimator. Let $S_n = \{(i,j) : \|X_i - X_j\| \leq h_1, i \neq j\}$,

$$\hat{\sigma^2} = \frac{1}{\#\{S_n\}}\sum_{(i,j)\in S_n}(Y_i - Y_j)^2.$$

It is easy to show that, if the regression function is Lipschitz continuous, as $h_1 \to 0, nh_1^p \to \infty$,

$$\hat{\sigma^2} - \sigma^2 = O(h_2^2) + O_p(\{nh_2^p\}^{-1/2}),$$

which implies that it is consistent.

A consistent estimator for the asymptotic variance for $\hat{\lambda}_i$ is thus given by

$$\hat{\sigma}_{ii}^L = \frac{\hat{\sigma}^2 J_2}{l^2\mu_2^2\hat{\delta}_i^2(l)}\sum_{k=1}^{l}\frac{1}{\hat{f}(\mathbf{x}_k)}\{\hat{U}_i^T(l)\hat{T}_{k+1}^l(1)\}^2\|\hat{T}^{k-1}\hat{V}_i(l)\|^2. \tag{5.39}$$

The following corollary follows from Theorem 5.5.

Corollary 5.5 *Under conditions of Theorem 5.5, and given consistent estimators* $\hat{\beta}_i$ *and* $\hat{\sigma}_{ii}^L$, *we have as* $h \to 0, nh^{p+2} \to \infty$,

$$(nh^{p+2})^{1/2}(\hat{\sigma}_{ii}^L)^{-1}(\hat{\lambda}_i - \lambda_i - \frac{h^2}{6\mu_2 l}\hat{\beta}_i - o(h^2)) \xrightarrow{d} N(0,1).$$

Based on Corollary 5.5, a pointwise confidence interval can be constructed for $\lambda_i(m)$ at point $\mathbf{x}_1$. Let

$$I_n = (\hat{\lambda}_i - (nh^{p+2})^{-1/2}\hat{\sigma}_{ii}^L Z_{\alpha/2} + \frac{h^2}{6\mu_2 l}\hat{\beta}_i, \hat{\lambda}_i + (nh^{p+2})^{-1/2}\hat{\sigma}_{ii}^L Z_{\alpha/2} + \frac{h^2}{6\mu_2 l}\hat{\beta}_i), \quad (5.40)$$

where $Z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal. By Corollary 5.5, if the remainder term in the asymptotic bias $o(h^2)$ is negligible, in the sense that $(nh^{p+2})^{1/2}h^2$ is bounded, i.e.

$$h \leq Cn^{-1/(p+6)}, \ C \text{ is a constant},$$

I_n is a confidence interval for $\lambda_i(l)$ with confidence level tending to $1-\alpha$, as $h \to 0, nh^{p+2} \to \infty$, i.e.

$$P(I_n) \to 1 - \alpha. \quad (5.41)$$

Simultaneous confidence intervals for several $\lambda_i(l)$'s can be constructed similarly.

In practice, the confidence interval (5.40) may still be too complicated to be computed. The main difficulty is the estimation of bias term $\hat{\beta}_i$ which requires the estimation of third-order partial derivatives. In nonparametric smoothing literature, similar problem occurs in constructing confidence interval for the regression function (Härdle 1990), and a simple way to avoid the problem is by choosing a smaller bandwidth, so that the bias term is negligible as compared with the variance, i.e. $O(h^2)$ is of smaller order than $O((nh^{p+2})^{1/2})$, then the bias adjustment is not needed. But this is done at the expense of increasing the variance, hence increasing the length of the confidence interval. That is, the confidence interval so constructed will be less precise than that using the bias adjustment.

Chapter 6

OUTLOOK

Current nonparametric functional estimation theory presumes the existence of a density function. The density function, when it exists, is certainly the most informative and convenient way to deal with. However, as shown in Section 4.5, if the probability measure is singular with the Lebesgue measure, the statistics can be drastically different from that assuming the existence of a density.

The immediate extension of the present work is to the case of a chaotic time series which does not possess a joint density function. The nonparametric estimation in fractal time series discussed in Section 4.5 can certainly be followed further. For example, the local linear fit and the local quadratic fit may be extended to fractal time series. A general local polynomial fit in fractal time series is given in Conjecture 4.2. Another issue is the phenomenon of *lacunarity* in chaotic time series (Smith 1991, 1992a). As pointed out by Richard Smith, the assumption in (4.9) is not general enough and does not cover the lacunar case, which turns out to be quite common. This issue is certainly worth looking into in the future.

The methods developed here will enable us to estimate the local Lyapunov exponents along with the estimates of variabilities of the estimators or confidence intervals. The logical next step is to test the methods with some known chaotic systems, and to apply them in analyzing some real data.

Moreover, the results on the local Lyapunov exponents may give some insights on the closely related problem of estimating the (global) Lyapunov exponents. An unsettled issue is the convergence rate of the Lyapunov exponent estimators, and the choice of block size l. See McCaffrey et al (1992), Ellner et al (1991) for more discussions on this open issue.

An important problem is nonlinear prediction. The nonparametric regression methods discussed in the present work are certainly very relevant. One issue is how to form multistep forecasts. See Farmer and Sidorowich (1988) for more discussions.

Bibliography

[1] Abarbanel, H.D.I., Brown, R. and Kennel, M.B. (1991). Lyapunov exponents in chaotic systems: their importance and their evaluation using observed data. *International J. Modern Physics B*, **5**, No. 9, 1347-1375.

[2] Bailey, B. (1993). Local Lyapunov exponents: predictability depends on where you are. Preprint, Department of Statistics, North Carolina State University.

[3] Barnsley, M.S. (1988). *Fractals Everywhere.* Boston: Academic Press.

[4] Bartlett, M. S. (1990). Chance or chaos? *J. R. Statist. Soc. A*, **153**, part 3, 321-347.

[5] Berliner, L.M. (1992). Statistics, probability and chaos. *Statist. Sci.*, **7**, No. 1, 69-122.

[6] Casdagli, M. (1992). Chaos and deterministic versus stochastic non-linear modeling. *J. R. Statist. Soc. B*, **54**, No. 2, 303-328.

[7] Castellana, J.V. and Leadbetter, M.R. (1986). On smoothed probability density estimation for stationary processes. *Stochastic Processes and their Applications*, **21**, 179-193.

[8] Chan, K.S. and Tong, H. (1994). A note on noisy chaos. *J. R. Statist. Soc. B*, **56**, No. 2, 301-311.

[9] Chatterjee, S and Yilmaz, M. (1992). Chaos, fractals and statistics. *Statist. Sci.*, **7**, No. 1, 49-122.

[10] Chu, C.K. and Marron, J.S. (1992). Choosing a kernel estimator. *Statist. Sci.,* **6**, No. 4, 404-425.

[11] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.,* **74**, 829-836.

[12] Cleveland, W. and Devlin, S. (1988). Locally weighted regression : an approach to regression analysis by local fitting. *J. Amer. statist. Assoc.,* **83**, 596-610.

[13] Devaney, R.L. (1989). *An Introduction to Chaotic Dynamical Systems,* second edition. Addison-Wesley Publishing Company.

[14] Eaton, M.L. and Tyler, D.E. (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random matrix. *Ann. Statist.*, **19**, No. 1, 260-271.

[15] Eckmann, J.-P. and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.,* **57**, 617-656.

[16] Ellner, S., Gallant, A.R., McCaffrey, D. and Nychka, D. (1991). Convergence rates and data requirements for Jacobian-based estimates of Lyapunov exponents from data. *Phys. Lett.* A, **153**, 357-363.

[17] Falconer, K. (1990). *Fractal Geometry.* Chichester: Wiley.

[18] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. of Statist.*, **21**, No. 1, 196-216.

[19] Fan, J. and Gijbels, I. (1992). Variable bandwith and local linear regression smoothers. *Ann. of Statist.*, **20**, 2008-2036.

[20] Farmer, J.D. and Ott, E. and Yorke, J.A. (1983). The dimension of chaotic attractors. *Physica D,* **7**, 153-180.

[21] Farmer, J.D. and Sidorowich, J.J.(1987). Predicting chaotic time series. *Physical Review Letters*, **59**, No.8, 845-848.

[22] Farmer, J.D. and Sidorowich, J.J.(1988). Exploiting chaos to predict the future and reduce Noise. *Evolution, Learning and Cognition.* Ed. Y. C. Lee. World Scientific Press. P277.

[23] Guckenheimer, J. and Holmes, P. (1990). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, third printing, revised and corrected. App. Math. Sci. **42**. New York: Springer.

[24] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Boston.

[25] Kifer, Y. (1986). *Ergodic Theory of Random Transformations.* Bikhäuser, Boston.

[26] Jazwinski, A. H.(1970). *Stochastic Processes and Filtering Theory.* Academic Press, New York.

[27] Lasota, A. and Mackey, M. C. (1994). *Chaos, Fractals and Noise. Stochastic Aspects of Dynamics.* Second Edition. App. Math. Sci. **97**. Springer-Verlag, New York.

[28] Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmospheric Science*, **20**, 130-141.

[29] Mandelbrot, B. (1982). *The fractal Geometry of Nature.* Freedman, San Francisco.

[30] Masry, E. and Fan, J. (1993). Local polynomial estimation of regression for mixing processes. *Mimeo Series (# 2311)*, Department of Statistics, University of North Carolina at Chapel Hill.

[31] McCaffrey, D., Nychka, D., Ellner, S. and Gallant, A.R. (1992). Estimating Lyapunov exponents with nonparametric regression. *J. Amer. Statist. Soc.* **87**, No.419, 682-695.

[32] Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data,* Springer Verlag, Berlin.

[33] Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators.* Cambridge: Cambridge University Press.

[34] Nychka,D., Ellner, S., McCaffrey, D. and Gallant, A.R. (1992). Finding chaos in noisy systems. *J. R. Statist. Soc. B*, **54**, No. 2, 399-426.

[35] Oseledec, V.I. (1968). A multiplicative ergodic theorem: Liapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.* **19**, 197-231. Reprinted in *Dynamical Systems*, 1991, edited by Ya. G. Sinai. Singapore: World Scientific.

[36] Ott, E. (1993). *Chaos in Dynamical Systems.* Cambridge University Press.

[37] Parker, T. S. and Chua, L. O. (1989). *Practical Numerical Algorithms for Chaotic Systems.* New York, Springer-Verlag.

[38] Pesin, Ya. B. and Sinai, Ya. G. (1981). Hyperbolicity and Stochasticity of Dynamical Systems. *Soviet Scientific Reviews. Section C, Math. Physics Reviews.* Gordon and Breach Press, Harwood Acad. Publ.. Vol. **2**, 1981. 53-115.

[39] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications,* second edition. New York: John Wiley & Sons.

[40] Rosenblatt, M. (1991). *Stochastic Curve Estimation.* NSF-CBMS Conference Series in Probability and Statistics, Volume 3. Institute of Mathematical Statistics.

[41] Ruelle, D. (1989). *Chaotic Evolution and Strange Attractors.* Cambridge: Cambridge University Press.

[42] Ruelle, D. (1990). Deterministic chaos: the science and the fiction. *Proc. R. Soc. Lond.* A,**427**, 241-248.

[43] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Technical Report (92-4)*, Department of Statistics, Rice University. (To appear in Annals of Statistics, 1994?)

[44] Sauer, T., Yorke, J.A., Casdagli, M. (1991). Embedology. *Journal of Statistical Physics.* Vol. 65, No. 3/4, 579-616.

[45] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons.

[46] Shiryayev, A.N. (1984). *Probability.* Graduate Texts in Mathematics: 95. Springer-Verlag.

[47] Sinai, Ya. G. (Ed)(1989). *Dynamical Systems II.* Springer-Verlag. Encyclopedia of Mathematical Sciences. Vol.2.

[48] Smith, R.L. (1991). Optimal estimation of fractal dimension. In *Nonlinear Modeling and Forecasting*, SFI Studies in the Sciences of Complexity, Proc. Vol. XII, Eds. M. Casdagli and S. Eubank, Addison-Wesley.

[49] Smith, R.L. (1992a). Estimating dimension in noisy chaotic time series. *J. R. Statist. Soc. B*, **54**, No. 2, 329-351.

[50] Smith, R.L. (1992b). Some comments on the relation between statistics and chaos. A discussion on two papers on statistics and chaos theory by Berliner, Chatterjee and Yilmaz. *Statist. Sci.*, **7**, No.1, 109-113.

[51] Sparrow, C. T. (1982). *The Lorenz Equations: Bifurcations, Chaos and Strange Attractors.* App. Math. Sci. **41**. Springer, New York.

[52] Stone, C.J.(1977). Consistent nonparametric regression. *Ann. of Statist.*, **5**, No.4, 595-645.

[53] Stone, C.J.(1980). Optimal rate of convergence for nonparametric estimators. *Ann. of Statist.*, **8**, No.6, 1348-1360.

[54] Sugihara, G. and May, R.M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, **344**, 734-741.

[55] Takens, F.(1981). Detecting strange attractors in turbulence. In Lecture Notes in Mathematics. No. 898, 366-381. Springer-Verlag.

[56] Tong, H. (1990). *Nonlinear Time Series: a Dynamical System Approach.* Oxford: Oxford University Press.

[57] Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D*, **16**, 285-317.

[58] Wolff, R.C.L. (1992). Local Lyapunov exponents: looking closely at chaos. *J. R. Statist. Soc. B*, **54**, No. 2, 353-371.