2001

Kernel smoothing for spatially correlated data

Xiao-Hu Liu
Iowa State University

Kernel smoothing for spatially correlated data

by

Xiao-Hu Liu

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Major Professors: Jean Opsomer and Kenneth Koehler

Iowa State University
Ames, Iowa
2001

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Kernel smoothing for spatially correlated data

Xiao-Hu Liu

Major Professors: Jean Opsomer and Kenneth Koehler
Iowa State University

Kernel smoothing is a nonparametric approach for estimating the relationship between a response variable and a set of predictors (or design variables). A major problem for kernel smoothing is the selection of the bandwidth, which controls the amount of smoothing. When data are correlated, former studies on kernel smoothing have been essentially limited to the case of a univariate predictor, with equally spaced design. In this dissertation, we discuss a more general case for correlated data, the case of multivariate predictors with random design. Three types of estimators, the Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator, are addressed, with emphasis on the local linear estimator. We will derive formulas for asympt-otic mean squared errors of these kernel smoothing estimators, and formulas of asymptotically optimal bandwidth. In the presence of spatially correlated errors, we show that traditional data-driven bandwidth selection methods, such as cross-validation and generalized cross-validation, fail to provide good bandwidth values. We propose several data-driven bandwidth selection methods that account for the presence of spatial correlation. Simulation studies show that these methods are effective when the covariances between the errors are completely known. When the covariances need to be estimated from data, we consider two special cases: spatial data with repeated measurements, and spatial data collected on a grid (with only one realization). For data with repeated measurements, we propose an estimation method based on semi-variogram fitting. For data on a grid, we propose a method based on differencing, with the application of approximate Whittle likelihood estimation. Simulation studies show that these methods can provide reasonably good estimates of the covariances for the purpose of bandwidth selection.

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Xiao-Hu Liu

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

In regression analysis, researchers are often interested in estimating the mean function $\mathrm{E}(Y|\boldsymbol{X}) = m(\boldsymbol{X})$ for a given set of observations $(\boldsymbol{X}_1, Y_1)$, $(\boldsymbol{X}_2, Y_2)$, ..., $(\boldsymbol{X}_n, Y_n)$, where the responses Y_i are scalar and the predictors (or design variables) $\boldsymbol{X}_i$ are either univariate or multivariate. Nonparametric regression is attractive since it does not require a parametric form for the mean function. Because of recent theoretical developments and widespread use of fast and inexpensive computers, nonparametric regression has become a rapidly growing and exciting field of statistics. Researchers have realized that for many real data sets, parametric regression is not sufficiently flexible to adequately fit curves or surfaces. Recent monographs on nonparametric regression (Eubank (1988), Müller (1988), Härdle (1990), Hastie and Tibshirani (1990), Wahba (1990), and Fan and Gijbels (1992a)) have shown that for a large variety of interesting examples, applications of nonparametric regression have yielded analyses essentially unobtainable by other techniques.

Roughly speaking, nonparametric regression techniques consist of basic smoothing methods and dimension reduction methods. Existing smoothing methods may be classified in three types: kernel smoothing methods, spline methods, and series expansion methods. The Priestley-Chao estimator, the Nadaraya-Watson estimator, the Gasser-Müller estimator, and the local polynomial estimator are all kernel smoothing methods. Spline methods include regression splines, smoothing splines, and penalized splines. Series expansion methods include polynomial series expansion, Fourier series expansion, and the wavelet method.

There have been some important and powerful dimension reduction methods which address the case of multivariate predictors, for instance, projection pursuit (Friedman and Stuetzle (1981)), ACE (Breiman and Friedman (1985)), generalized additive models (Hastie and Tibshirani (1990)), and MARS (Friedman (1991)). These techniques have been implemented in some commercial software packages and are widely used as data mining tools.

The focus of this dissertation will be kernel smoothing methods for correlated data. The bulk of literature in kernel smoothing has focused on the situation of uncorrelated data, while limited work has addressed the situation of correlated data. We will look at the following model:

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i \quad (i = 1, \ldots, n) \tag{1.1}$$

where $m(\cdot)$ is an unknown, smooth mean function that needs to be estimated, the ε_i are random errors, the $\boldsymbol{X}_i$ are either random or fixed with domain Ω, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_1)^T$ has zero mean and the covariance matrix $\sigma^2\boldsymbol{\rho}$. The correlation matrix $\boldsymbol{\rho}$ is either considered completely known, known up to a finite number of parameters, or left completely unspecified. We will use $f(\boldsymbol{x})$ to denote the probability density function of the design variable $\boldsymbol{X}_i$.

As a convention, we will use lower case letter $\boldsymbol{x}$ when the design is fixed and capital letter $\boldsymbol{X}$ when it is random. Generally, bold letters will be reserved for vectors and matrices. For a vector $\boldsymbol{u} = (u_1, \ldots, u_p)^T$ and an integrable function $q(\cdot)$, the multiple integral $\int\int\cdots\int q(\boldsymbol{u})du_1du_2\cdots du_p$ will be simply denoted as $\int q(\boldsymbol{u})d\boldsymbol{u}$. For any matrix $\boldsymbol{B}$, we use $|\boldsymbol{B}|$, $\text{tr}(\boldsymbol{B})$, $\lambda_{\max}(\boldsymbol{B})$, $\lambda_{\min}(\boldsymbol{B})$, and $||\boldsymbol{B}||$, to denote its determinant, trace, maximum eigenvalue, minimum eigenvalue, and L_2 norm, respectively. Note that $||\boldsymbol{B}|| = \sqrt{\lambda_{\max}(\boldsymbol{B}^T\boldsymbol{B})}$. If $\boldsymbol{B}$ is symmetric, the notation $\boldsymbol{B} \geq \boldsymbol{0}$ ($> \boldsymbol{0}$) means that $\boldsymbol{B}$ is nonnegative definite (positive definite). If $\boldsymbol{A}$ and $\boldsymbol{B}$ both are symmetric, nonnegative definite, we may use $\boldsymbol{A} \leq \boldsymbol{B}$ to mean that $\boldsymbol{B} - \boldsymbol{A} \geq \boldsymbol{0}$, or $\boldsymbol{B} - \boldsymbol{A}$ is nonnegative definite.

This chapter gives an overview of the methods of kernel smoothing. Section 1.1 introduces the basic concepts of univariate kernel smoothing, and various univariate kernel smoothing estimators. Section 1.2 discusses the case of multivariate predictors. Section 1.3 summarizes recent developments in kernel smoothing methods with correlated data, particularly for the univariate case with equally spaced designs. Section 1.4 is the outline of this dissertation.

1.1 Univariate kernel smoothing

1.1.1 Basic concepts

In this section we assume that the design variable or predictor is univariate. Without specifying the parametric form of the regression function m, it is reasonable to assume that a design point far from x carries relatively little information about the value of $m(x)$. Hence an intuitive estimator for $m(x)$ is the running local average. A better version of this is the locally weighted average. The weights are determined by a nonnegative function $K(\cdot)$, called the kernel function, and by a positive number h, called the bandwidth. Usually the kernel is a symmetric probability density function. For theoretical simplicity, we assume that the support of the kernel is $(-1, 1)$. The bandwidth h is used to control the size of the neighborhood of x, where the weights have nonzero values. The parameter h is referred to as the smoothing parameter, because it controls the amount of smoothing.

Let $K_h(u) = K(u/h)/h$. Then the support of $K_h(u)$ is $(-h, h)$. For a given point x, the weight of a datum (X_i, Y_i) is $K_h(X_i - x)$. It has a nonzero value only if $X_i \in (x - h, x + h)$.

Examples of commonly used kernels include:

- the uniform kernel

$$K(u) = \frac{1}{2} I_{(-1,1)}(u).$$

- the quadratic (Epanechnikov) kernel

$$K(u) = \frac{3}{4}(1-u^2) I_{(-1,1)}(u).$$

- the biweight kernel

$$K(u) = \frac{15}{16}(1-u^2)^2 I_{(-1,1)}(u).$$

These kernels are all of second order, in the sense that their first moment $\int_{-1}^{1} uk(u)du = 0$, and their second moment $\int_{-1}^{1} u^2 k(u)du$ is finite. We are only interested in second order kernels. Sometimes, however, kernels with an order higher than 2 are used in applications. A kernel $K(u)$ is called a kernel of order p if the first $p-1$ moments of $K(u)$ are 0 and the p-th moment is finite. Note that $k(u)$ has to be negative somewhere if its order is higher than 2.

Let $\hat{m}(x,h)$ represent the kernel estimator for the value of the mean function m at x using bandwidth h. This estimator could be obtained from locally weighted average estimation or local polynomial estimation (defined later in this section). The performance of this estimator is assessed by its *Mean Squared Error* (MSE):

$$\text{MSE}(x,h) = \text{E}\left\{(\hat{m}(x,h) - m(x))^2 | X_1, \ldots, X_n\right\}, \tag{1.2}$$

or by its Mean Integrated Squared Error:

$$\text{MISE}(h) = \int \text{MSE}(x,h) w(x) dx, \tag{1.3}$$

with $w(x) \geq 0$, a weight function specified by the user. The MSE criterion is used to assess the performance of the estimator at a given point x, while the MISE is used to assess the performance of the estimator in recovering the whole curve. The MSE has the following bias-variance decomposition

$$\text{MSE}(x,h) = (\text{E}\{\hat{m}(x,h)|X_1, \ldots, X_n\} - m(x))^2 + \text{Var}\{\hat{m}(x,h)|X_1, \ldots, X_n\}, \tag{1.4}$$

where we refer to the term $\mathrm{E}\{\hat{m}(x,h)|X_1,\ldots,X_n\}-m(x)$ as the (conditional) bias of the estimator $\hat{m}(x,h)$ and refer to the term $\mathrm{Var}\{\hat{m}(x,h)|X_1,\ldots,X_n\}$ as the (conditional) variance of $\hat{m}(x,h)$. The large sample approximations of MSE and MISE will be important for selecting a bandwidth by a data-driven method, especially a plug-in type method. They are denoted as AMSE and AMISE respectively, and will be discussed later in this chapter.

1.1.2 Local weighted average estimator

We will consider three types of local weighted average estimators: the Nadaraya-Watson estimator (Nadaraya (1964), and Watson (1964)), the Priestley-Chao estimator (Priestley and Chao (1972)), and the Gasser-Müller estimator (Gasser and Müller (1979)).

The Nadaraya-Watson estimator is defined as

$$\hat{m}(x,h) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{1.5}$$

it can be used with non-uniformly random designs, but the random denominator in its expression is inconvenient when deriving its asymptotic properties. This motivates the introduction of the Priestley-Chao estimator and the Gasser-Müller estimator.

Assume that the data have been sorted according to the X-variable. The Priestley-Chao estimator was defined as

$$\hat{m}(x,h) = \sum_{i=1}^{n}(X_i - X_{i-1})K_h(X_i - x)Y_i \ , \tag{1.6}$$

with $X_0 = 0$. The Priestley-Chao estimator is difficult to extend to the higher dimensional case because its calculation involves sorting of the design points, which is not computationally trivial in higher dimensional spaces.

For equally spaced fixed designs and uniformly random designs in the region $\Omega =$

$[0, b]$, sorting of the design points is not necessary. Later in this dissertation, we will call

$$\hat{m}(x,h) = \frac{b}{n}\sum_{i=1}^{n} K_h(X_i - x)Y_i \tag{1.7}$$

the Priestley-Chao estimator. This estimator has a very simple expression and can be applied to higher dimensional problems. But it is asymptotically biased in the non-uniform random design case.

Again, we assume that the data have been sorted according to the X-variable. The Gasser-Müller estimator is defined as

$$\hat{m}(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(u - x)du Y_i, \tag{1.8}$$

with $s_i = (X_i + X_{i+1})/2$, $X_0 = -\infty$, and $X_{n+1} = +\infty$. The Gasser-Müller estimator is also difficult to extend to the higher dimensional case because its calculation involves sorting and taking middle points in the design space.

1.1.3 Local polynomial regression

Local polynomial regression was introduced by Stone (1977). It was systematically studied by Stone (1977, 1980, 1982), Cleveland (1979), Fan (1992,1993), Fan and Gijbels (1992b), and Ruppert and Wand (1994). These papers have clearly shown significant advantages of local polynomial regression over the local weighted average estimators when the mean function is smooth enough and the errors are uncorrelated.

The idea of local polynomial regression is simple. If the mean function $m(x)$ is smooth enough, then in a neighborhood of a given point x, we can well approximate $m(x)$ using a polynomial of p-th order. This suggests using a locally weighted polynomial regression by solving the weighted least squares problem

$$\min_{\{\beta_j\}_{j=0}^{p}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right)^2 K_h(X_i - x) \tag{1.9}$$

for β_j $(j = 0, \ldots, p)$. Let $\hat{\beta}_j$ $(j = 0, \ldots, p)$ denote the minimizer. Then the local polynomial regression estimator of $m(x)$ is

$$\hat{m}(x, h) = \hat{\beta}_0 \, . \tag{1.10}$$

A byproduct of local polynomial regression is the estimator of the derivatives of the mean function. That is, the estimator of the k-th order derivative $m^{(k)}(x)$ is simply $k!\hat{\beta}_k$ (for $k = 1, \ldots, p$).

We will focus on local linear regression ($p = 1$), because the local linear estimator possesses many of the asymptotically optimal properties of local polynomial regression (with an order higher than 1), and it is relatively easy to extend to the case of multivariate predictors, both in theory and calculations.

Let $b_n = \frac{1}{2}h^2 \int u^2 K(u)du$ and $V_n = \frac{\sigma^2}{f(x)nh} \int K^2(u)du$. Table 1.1 summarizes the asymptotic performance of the Nadaraya-Watson estimator, the Gasser-Müller estimator, and the local linear estimator for the mean function $m(x)$ when data are uncorrelated.

Table 1.1 Point-wise asymptotic bias and variance of kernel smoothers, taken from Fan (1992).

Estimator	Asymptotic Bias	Asymptotic Variance
Nadaraya-Watson	$\left(m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right) b_n$	V_n
Gasser-Müller	$m''(x)b_n$	$1.5V_n$
Local linear	$m''(x)b_n$	V_n

Compared to the local linear estimator, the Nadaraya-Watson estimator has larger bias, particularly in the region where both the derivative of the mean function and the ratio of the derivative of the design density to the design density itself are large. It may have bias even when the mean function is linear. Fan (1992, 1993) further showed that the Nadaraya-Watson estimator has zero minimax efficiency. The Gasser-Müller estimator corrects the bias of the Nadaraya-Watson estimator at the expense of increasing its variance. In addition, both the Nadaraya-Watson estimator and the Gasser-Müller estimator have a large order of bias when estimating a curve at a boundary region. The methods that deal with this issue, such as *reflection methods* and *boundary correction methods*, are less efficient than the automatic boundary correction of the local linear estimator. Cheng, Fan, and Marron (1993) showed that the local linear estimator is efficient in correcting boundary bias in an asymptotic minimax way.

For a uniform random design or fixed equally spaced design, the Priestley-Chao estimator has the same asymptotic bias and the same asymptotic variance as the local linear estimator. But as stated previously, it can not be adapted in an asymptotically unbiased way to multivariate predictors with a nonuniform random design.

1.2 Multivariate kernel smoothing

In this section we discuss the case of multivariate predictors. Suppose the the $\boldsymbol{X}_i$ are d-variate predictors and the design is random. The kernel $K(\boldsymbol{u})$ is d-variate and satisfies $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$. We assume that the kernel K is bounded, has support in a bounded region, and

$$\int \boldsymbol{u}\boldsymbol{u}^T K(\boldsymbol{u})d\boldsymbol{u} = \mu_2(K)\boldsymbol{I}_d,$$

where $\mu_2(K) \neq 0$ is scalar and $\boldsymbol{I}_d$ is the $d \times d$ identity matrix. Furthermore, we restrict our discussion to kernels with zero values for their odd-order moments. That is,

$$\int u_1^{l_1} \cdots u_d^{l_d} K(\boldsymbol{u})d\boldsymbol{u} = 0$$

for all non-negative integers $l_1, \ldots, l_d$ with their sum equal to an odd integer. The kernels that satisfy these conditions include spherically symmetric kernels and kernels that are a product of symmetric univariate kernels. In Chapter 2, for simplicity, we will assume the support of $K(\boldsymbol{u})$ to be $\{\boldsymbol{u} : ||\boldsymbol{u}|| \leq 1\}$.

In the multivariate case, the bandwidth $\boldsymbol{H}$ is now a matrix. The function

$$K_{\boldsymbol{H}}(\boldsymbol{u}) = |\boldsymbol{H}|^{-1} K\left(\boldsymbol{H}^{-1}\boldsymbol{u}\right)$$

is used to assign weights. $\boldsymbol{H}$ is a $d \times d$ symmetric, positive definite matrix depending on the sample size n. Given a point $\boldsymbol{x}$, the bandwidth $\boldsymbol{H}$ controls the shape and the size of the local neighborhood used for estimating $m(\boldsymbol{x})$. Only the design points within this neighborhood carry nonzero weights and are used for estimation.

To better understand how $\boldsymbol{H}$ determines the local neighborhood, let us consider the bivariate case ($d = 2$) when $K(\boldsymbol{x})$ is spherically symmetric with support $\{\boldsymbol{x} \in I\!R^2 : \|\boldsymbol{x}\| < 1\}$. $\boldsymbol{H}$ can be decomposed into

$$\boldsymbol{H} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}^T \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix},$$

with $\lambda_1 > 0$, $0 < \lambda_2 \leq \lambda_1$, and $0 < \theta \leq \pi$. The corresponding local neighborhood ("local bandwidth region") used to estimate $m(\boldsymbol{x})$ is

$$\{\boldsymbol{X} \in \Omega : \|\boldsymbol{H}^{-1}(\boldsymbol{X} - \boldsymbol{x})\| < 1\},$$

i.e.

$$\{\boldsymbol{X} \in \Omega : (\boldsymbol{X} - \boldsymbol{x})^T \boldsymbol{H}^{-2} (\boldsymbol{X} - \boldsymbol{x}) < 1\}$$

Geometrically, this neighborhood is an ellipse in $I\!R^2$ centered at $\boldsymbol{x}$, with the longer axis of length λ_1, the shorter axis of length λ_2, and with θ as the angle between the longer axis of the ellipse and the axis corresponding to the first design variable.

Next, we will define three types of kernel smoothing estimators for multivariate cases.

For a uniform random design, the Priestley-Chao estimator of $m(\boldsymbol{x})$ is defined as

$$\hat{m}(\boldsymbol{x}, \boldsymbol{H}) = \frac{b}{n} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) Y_i, \tag{1.11}$$

where b is the volume of the design region Ω. Note that the design region may be any bounded set.

The Nadaraya-Watson estimator is well defined in the multivariate case by

$$\hat{m}(\boldsymbol{x},\boldsymbol{H}) = \frac{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})Y_i}{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})}. \tag{1.12}$$

The local linear estimator will be our major focus in this dissertation. The estimator of m($\boldsymbol{x}$) is $\hat{\beta}_0$, where $\hat{\beta}_0$ combined with $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)^T$ is the solution to the following least squares problem:

$$\min_{(\beta_0, \boldsymbol{\beta}^T)} \sum_{i=1}^{n} \{Y_i - \beta_0 - \boldsymbol{\beta}^T(\boldsymbol{X}_i - \boldsymbol{x})\}^2 K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}). \tag{1.13}$$

We introduce the following matrix notation:

$$\boldsymbol{X_x} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^T \\ \vdots & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^T \end{pmatrix},$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T,$$

and

$$\boldsymbol{W_x} = \text{diag}(K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{x}), \ldots, K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{x})).$$

If the rank of the matrix $\boldsymbol{X_x}$ is equal to the number of the columns of $\boldsymbol{X_x}$, the solution to problem (1.13) is

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = (\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W_x} \boldsymbol{X_x})^{-1} \boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W_x} \boldsymbol{Y}. \tag{1.14}$$

The local linear estimator is

$$\hat{m}(\boldsymbol{x},\boldsymbol{H}) = \boldsymbol{e}_1^T (\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W_x} \boldsymbol{X_x})^{-1} \boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W_x} \boldsymbol{Y} \equiv \boldsymbol{s}_{\boldsymbol{x}}^T \boldsymbol{Y}, \tag{1.15}$$

with $\boldsymbol{e}_1$ as the $(d+1) \times 1$ vector having 1 in the first entry and 0 in all other entries.

For the case of uncorrelated data with a random design, Ruppert and Wand (1994) derived the asymptotic mean squared error (AMSE) formula for the multivariate local linear estimator. Let

$$\mu(K^2) \equiv \int K^2(\boldsymbol{x})d\boldsymbol{x},$$

and assume that the mean function $m(\boldsymbol{x})$ be second order differentiable, with $\mathcal{H}_m(\boldsymbol{x})$ as its Hessian matrix. For an interior point $\boldsymbol{x}$ in the support of the design density f, if $n \to \infty$, $\boldsymbol{H} \to 0$, $\frac{\lambda_{\max}(\boldsymbol{H})}{\lambda_{\max}(\boldsymbol{H})}$ is bounded above, and $n|\boldsymbol{H}|\lambda^2_{\min}(\boldsymbol{H}) \to \infty$, then

$$\text{AMSE}(\boldsymbol{x}, \boldsymbol{H}) = \left(\frac{1}{2}\mu_2(K)\,\text{tr}(\boldsymbol{H}\mathcal{H}_m(\boldsymbol{x}))\right)^2 + \frac{\mu(K^2)\sigma^2}{n|\boldsymbol{H}|f(\boldsymbol{x})}. \tag{1.16}$$

Here the first part of the right hand side is the squared bias, and the second part is the variance. Notice that the conditions which are imposed on $\boldsymbol{H}$ here are more restrictive than the corresponding conditions stated by Ruppert and Wand (1994). We need these conditions to complete the derivation of the above equation and other results. More details will be provided in Chapter 2.

1.3 Univariate kernel smoothing for correlated data

Opsomer *et al.* (2001) provide a good review of this topic. Some contents of this section are from their paper.

We consider the case where the design variable is univariate. We assume that the design points x_i are within a finite interval $[a, b]$, and for simplicity we assume $[a, b] = [0, 1]$. We consider the problem of how to estimate the mean function m for data assumed to follow model (1.1), with

$$\text{Var}(\varepsilon_i) = \sigma^2, \quad \text{Cor}(\varepsilon_i, \varepsilon_j) = \rho_n(x_i - x_j). \tag{1.17}$$

The dependence of the correlation ρ_n on the sample size is indicated by the subscript. The consistency properties of the estimators will depend on the behavior of the correlation function as n increases. Most previous researchers have focused on the time series case, where the design points are fixed and equally spaced, or equivalently,

$$x_i = \frac{i}{n},$$

and

$$Y_i = m\left(\frac{i}{n}\right) + \varepsilon_i, \quad \mathrm{Cor}(\varepsilon_i, \varepsilon_j) = \rho_n\left(\left|\frac{i}{n} - \frac{j}{n}\right|\right). \tag{1.18}$$

An important special case of the correlation function is

$$\rho_n(i/n) = \rho(i/n),$$

where $\rho(x)$ is continuous. In this case, the error process is a realization of a continuous process on $[0, 1]$. This process was discussed by Hart and Wehrly (1986) and Parzen (1959, 1961). They have shown that if only a single realization of the process has been observed, there is no consistent linear estimator for the mean function as the design points are sampled more and more densely on the unit interval.

Another important case is

$$\rho_n(i/n) = \rho(i),$$

where the error process is constant, no matter how close together the design points become. In other words, the correlation between two points is only related to the difference of the indices of the two points. We will consider this simplest situation in depth below.

Altman (1990) considered the Priestley-Chao estimator on the interval [0, 1],

$$\hat{m}(x, h) = \boldsymbol{s}_x^T \boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)Y_i\ . \tag{1.19}$$

The mean squared error of $\hat{m}(x, h)$ is given by

$$\mathrm{MSE}(\hat{m}(x, h)) = \mathrm{E}(\boldsymbol{s}_x^T \boldsymbol{Y} - m(x))^2 = (\boldsymbol{s}_x^T \mathrm{E}(\boldsymbol{Y}) - m(x))^2 + \sigma^2 \boldsymbol{s}_x^T \boldsymbol{\rho} \boldsymbol{s}_x\ . \tag{1.20}$$

The first term of the MSE represents the squared bias, while the second term corresponds to the variance. Under the usual assumptions on the kernel and the bandwidth, and for a kernel of order p, the bias is approximated by

$$\boldsymbol{s}_x^T \mathrm{E}(\boldsymbol{Y}) - m(x) = (-h)^p \frac{\mu_p(K)}{p!} m^p(x) + o(h^p) \tag{1.21}$$

(see Altman (1990)), with $\mu_q(K) = \int u^q K(u)du$ for any natural number q. Note the important fact that the bias of $\hat{m}(x,h)$ does not depend on the error correlation structure, which is also true for random designs and higher dimensional cases.

To study the variance part, we make some assumptions about the correlation function ρ:

(I) $\sum_{k=1}^{\infty} |\rho(k)| < \infty$,

(II) $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} k|\rho(k)| = 0$.

These assumptions, common in time series analysis, ensure that observations sufficiently far apart are essentially uncorrelated. Then, the variance component of the MSE can be approximated by

$$\sigma^2 \boldsymbol{s}_x^T \boldsymbol{\rho} \boldsymbol{s}_x = \frac{1}{nh}\mu(K^2)\sigma^2(1+2R) + o\left(\frac{1}{nh}\right) \tag{1.22}$$

with $R = \sum_{k=1}^{\infty} \rho(k)$ (Altman (1990)). When the observations are uncorrelated, this result reduces to the one usually reported for kernel regression with independent errors. Note also that the power spectral density of the errors is

$$S(\omega) = \sigma^2 \sum_{-\infty}^{\infty} \rho(|k|)e^{-ik\omega}, \tag{1.23}$$

so that

$$\sigma^2(1+2R) = S(0), \tag{1.24}$$

the spectral density at $\omega = 0$. This fact was used by Chiu (1989) for developing bandwidth selection methods based on frequency domain estimation.

When the kernel is of second order, the asymptotic approximation of the MSE is

$$\text{AMSE}(x,h) = \left(h^2\frac{\mu_2(K)}{2}m''(x)\right)^2 + \frac{1}{nh}\mu(K^2)\sigma^2(1+2R). \tag{1.25}$$

The presence of the additional term R in the AMSE has important implications for the correct choice of the bandwidth. If $R > 0$, implying the error correlation is positive in total, then the variance of $\hat{m}(x, h)$ will be larger than that in the uncorrelated case. Hence the AMSE is minimized by a larger bandwidth value h compared to the uncorrelated case. If $R < 0$, the AMSE-optimal bandwidth is smaller than that in the uncorrelated case.

In addition to changing the asymptotically optimal bandwidth, correlation has a perverse effect on automated bandwidth selection methods as well, as described in Altman (1990) and Hart (1991) for cross-validation. As a global measure of goodness-of-fit, we consider the Mean Average Squared Error (MASE), or

$$\text{MASE}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}(\frac{i}{n}) - m(\frac{i}{n})\right)^2 . \tag{1.26}$$

Let $\hat{m}_{(-i)}$ denote the kernel regression estimate computed on the data set with the i-th observation not used. The cross-validation criterion is

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_{(-i)}(\frac{i}{n}) - Y_i\right)^2 . \tag{1.27}$$

For the Priestley-Chao estimator and Gasser-Müller estimator, Altman (1990) and Hart (1991) showed that asymptotically the CV criterion tends to select small bandwidth h that leads to a fit which nearly interpolates the data. The GCV criterion and Mallows C_L criterion also break down for correlated data.

One approach to fix this problem is to estimate the correlation function parametrically and to use this estimate to adjust the bandwidth selection criterion. The estimation of the correlation function is complicated by the fact that the errors are unobserved. Chiu (1989) attempted to bypass this problem by estimating the correlation function in the frequency domain while down-weighting the low frequency periodogram components where the slowly varying mean function usually lie. Altman (1990) used residuals from a kernel regression fit using a pilot bandwidth and fitted a low-order autoregressive process to the residuals. The major difficulty for Altman's method is how to choose the pilot bandwidth. Indeed, when the pilot bandwidth is too small, the smoothing curve tends to interpolate the data. Thus, the correlation of the residuals tends to be too small compared to the correlation of the errors. In contrast, when the pilot bandwidth is too large, the bias will usually be large and may greatly distort the residuals away from the true errors. This may lead to very inaccurate estimation of the true correlation function. Hart (1991) considered the Gasser-Müller estimator. He used differencing to remove the trend followed by estimation of the correlation function using a spectral density and Whittle likelihood. This method is very appealing because it does not require a pilot fit, but it is hard to apply to random designs or fixed unequally spaced designs. In his later work, Hart (1994) introduced the method of *time series cross-validation* as a goodness-of-fit criterion, which can be jointly minimized over the set of parameters for the correlation function and the bandwidth parameter. All the above methods appear to work well in simulation studies. Even when the parametric form of the correlation structure is misspecified, they provide a significant improvement over the fits computed under the assumption of uncorrelated errors. However the problem of estimating the correlation function is far from solved and extremely important.

More recently, Masry (1995, 1996), and Masry and Fan (1997) discussed the local polynomial regression estimator for strongly mixing and ρ-mixing time series processes. They gave a result in which the asymptotic variance of the estimator of the mean function does not depend on the correlations between errors. Francisco-Fernández and Vilar-Fernández (2001) discussed the local polynomial regression estimator for correlated data under a "regular" fixed design, where the design points are generated from a design density function. They proposed a plug-in type method for bandwidth selection, starting with a pilot bandwidth computed by Hart's time series cross-validation method (see Hart (1994)). These studies are restricted to the univariate, nonrandom designs. Opsomer (1997) initiated the research on local linear regression for correlated data in the case of multivariate predictor under random designs, but did not provide a method for bandwidth selection.

Another approach is completely nonparametric. It avoids specifying a parametric model for the correlation function. This approach will be briefly introduced without further detailed discussion.

Chu and Marron (1991) proposed two new cross-validation based criteria that estimate the optimal bandwidth that minimizes the mean average squared errors without specifying the correlation function. In modified cross-validation (MCV), the kernel regression values $\hat{m}_{(-i)}$ in (1.27) is substituted by a result that is computed by leaving out the $2l+1$ observations $i-l, i-l+1, \ldots, i+l-1, i+l$ surrounding the ith observation. In partitioned cross-validation (PCV), the observations are partitioned into g subgroups by taking every g-th observation. Within each subgroup, the observations are further apart and, hence, are less correlated. Cross-validation is performed for each subgroup, and the bandwidth estimate for all the observations is a simple function of the average of the subgroup-optimal bandwidths. The drawback of both MCV and PCV is that the values of l and g need to be selected with some care.

Herrmann *et al.* (1992) also proposed a fully nonparametric method for estimating the MASE-optimal bandwidth, but replaced the CV-based criterion by a *plug-in* approach. It is easy to show that the minimizer of the MASE is asymptotically equal to

$$h_{\mathrm{MASE}} \approx c(K)\left(\frac{\sigma^2(1+2R)}{n\int m''(x)^2dx}\right)^{1/5},$$

with $c(K)$ as a known kernel-dependent constant. Plug-in bandwidth selection is performed by estimating the unknown quantities in this expression and replacing them by their estimators. The estimation of $\int m''(x)^2dx$ is completely analogous to that in the uncorrelated case. The variance component $\sigma^2(1+2R)$ is estimated by a summation over squared differencing residuals. As in Chu and Marron (1991), no parametric form is assumed for the correlation function.

Hall *et al.* (1995) extended the results of Chu and Marron (1991) in a number of useful directions. Their theoretical results apply to kernel regression as well as local linear regression. They also explicitly considered the long-range dependence case. where assumptions (I) and (II) are no longer required. They discussed bandwidth selection through MCV and compared it with a bootstrap-based approach which estimates the MASE in (1.26) directly through resampling of "blocks" of residuals from a pilot fit. Successful application, however, requires a careful choice of other tuning parameters.

1.4 Dissertation organization

As discussed in the above section, for correlated data, research on kernel smoothing methods has been essentially restricted to the univariate case, with fixed and equally spaced designs. In this dissertation, we will look at the multivariate case with random designs. In chapter 2, we will derive the formulas for the asymptotic mean squared errors of three types of kernel smoothing estimators (the Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator). The asymptotic optimal bandwidth formulas will also be derived. In Chapter 3, we will discuss the selection of the bandwidth matrix $\boldsymbol{H}$ in the case where the covariance matrix $\sigma^2\boldsymbol{\rho}$ is completely known. We will show how, in the presence of spatially correlated errors, traditional data-driven bandwidth selection methods fail to provide good bandwidth values. We will propose some new bandwidth selection criteria that need the information of covariances of the errors. In chapter 4, we will address the issue of bandwidth selection when the covariance matrix $\sigma^2\boldsymbol{\rho}$ is estimated from the data. We will consider two kinds of situations : spatial data with repeated measurements, and spatial data collected on a grid. For spatial data with repeated measurements, semi-variogram fitting will be used for estimating covariances between the errors. For spatial data on a grid, differencing and approximate bivariate Whittle likelihood will be used. By simulation study, we will show that the estimates of the covariances from these methods are reasonably good for bandwidth selection. General conclusions will be presented in Chapter 5.

2 ASYMPTOTIC PROPERTIES OF KERNEL SMOOTHING ESTIMATORS WITH CORRELATED DATA

2.1 Introduction

In this chapter, we discuss the asymptotic properties of multivariate kernel smoothing estimators for correlated data under random designs. We will consider three types of estimators: the Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator. We will derive formulas for asymptotic mean squared errors and formulas for asymptotic optimal bandwidths for these estimators.

When data are uncorrelated, the asymptotic mean squared error of the multivariate local linear estimator was studied by Ruppert and Wand (1994). They also considered the local quadratic estimator, which will not be pursued in this dissertation. When data are correlated, the problem becomes more difficult. Most authors have focused on the univariate case, in time series settings or with fixed and equally spaced designs. Altman (1990) has considered the Priestley-Chao estimator. Hart (1991) has examined the Gasser-Müller estimator. The univariate local polynomial estimator has been studied by Masry (1995, 1996), Masry and Fan (1997), Francisco-Fernández and Vilar-Fernández (2001). The research on multivariate local linear regression with correlated data was initiated by Opsomer (1997), who addressed the random design case.

We will consider the multivariate version of the Priestley-Chao estimator for uniformly random designs. For a general random design, we will consider the Nadaraya-Watson estimator and the local linear estimator. As discussed in Chapter 1, the local linear estimator has better asymptotic properties than the local weighted average estimators in the independent error case. Hence it will be our major focus.

We consider the model:

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i, \tag{2.1}$$

where Y_i are scalar, and the covariates (design points) $\boldsymbol{X}_i \in \Omega$ are random ($i = 1, 2, \ldots .n$). Here Ω is a fixed domain, a bounded open closure in $I\!R^d$ (A set is referred to as an open closure if the set is a union of an open set and the boundary of this open set). Let $f(\boldsymbol{x})$, the density function of the covariate $\boldsymbol{X}$, have continuous second-order derivatives, with gradient vector $\nabla_f(\boldsymbol{x})$ and Hessian matrix $\boldsymbol{\mathcal{H}}_f(\boldsymbol{x})$ (with second order partial derivatives as its entries). Let $m(\boldsymbol{x})$ be second order differentiable, with Hessian matrix $\boldsymbol{\mathcal{H}}_m(\boldsymbol{x})$. Let

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j | \boldsymbol{X}_i, \boldsymbol{X}_j) = \sigma^2 \rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j), \tag{2.2}$$

where $\rho_n(\boldsymbol{x})$ is continuous, satisfying $\rho_n(\mathbf{0}) = 1$, $\rho_n(\boldsymbol{x}) = \rho_n(-\boldsymbol{x})$, and $|\rho_n(\boldsymbol{x})| \leq 1 \;\forall\; \boldsymbol{x}$. In case of a random design, covariances between random errors are related to the location of the design points, which are random. Notice that the correlation function ρ_n is also related to the sample size n. To ensure the consistency of the kernel smoothing estimators, ρ_n has to "shrink" as the sample size $n \to \infty$ (see below). Let the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ have zero mean and variance matrix

$$\boldsymbol{V} \equiv \mathrm{Var}(\boldsymbol{\varepsilon} | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \sigma^2 \boldsymbol{\rho} = \sigma^2 \left(\rho_n(\mathbf{X}_i - \boldsymbol{X}_j)\right)_{n \times n}.$$

Define the response vector

$$\boldsymbol{Y} = (y_1, \ldots, y_n)^T,$$

and the mean vector

$$\boldsymbol{m} = (m(\boldsymbol{X}_1), \ldots, m(\boldsymbol{X}_n))^T.$$

Let $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ represent the kernel smoothing estimator of the mean function $m(\boldsymbol{x})$ obtained by using bandwidth $\boldsymbol{H}$. For simplicity, the kernel $K(\boldsymbol{x})$ is assumed to be a spherically symmetric density function, with a nonzero value only if $\|\boldsymbol{x}\| < 1$. But the results in this chapter may hold for more general kernels, for instance, a kernel that is a tensor product of univariate symmetric kernels.

The following are two examples of spherically symmetric d-dimensional kernels:

- Uniform kernel

$$K(\boldsymbol{x}) = \begin{cases} 1/v_d & \text{if } \|\boldsymbol{x}\| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

  where v_d is the volume of the d-dimensional unit ball $\{\boldsymbol{x} : \boldsymbol{x} \in I\!R^d, \|\boldsymbol{x}\| < 1\}$.

- Epanechnikov kernel

$$K(\boldsymbol{x}) = \begin{cases} \frac{d(d+2)}{2S_d}(1 - \|\boldsymbol{x}\|^2)\, 1_{(\|\boldsymbol{x}\|<1)} & \text{if } \|\boldsymbol{x}\| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

  where $S_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the area of the surface of the d-dimensional unit ball $\{\boldsymbol{x} : \boldsymbol{x} \in I\!R^d, \|\boldsymbol{x}\| < 1\}$.

Following notation introduced in Chapter 1,

$$\mu(K^2) \equiv \int K^2(\boldsymbol{x})d\boldsymbol{x}$$

and

$$\int K(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T d\boldsymbol{x} = \mu_2(K)\boldsymbol{I}_d.$$

Also, we will use $O(\cdot)$ and $o(\cdot)$ to represent the convergence rate of a sequence, $O_p(\cdot)$ and $o_p(\cdot)$ to represent the convergence (in probability) rate of a random sequence. If a random sequence Z_n converges to 0 in probability, then $Z_n = o_p(1)$. If $\boldsymbol{U}_n$ is a random matrix, then $O_p(\boldsymbol{U}_n)$ and $o_p(\boldsymbol{U}_n)$ are to be taken component-wise.

Starting from here, throughout this dissertation, we assume that the above regularity conditions stated for the design region Ω, the mean function $m(\boldsymbol{x})$, the design density $f(\boldsymbol{x})$, and the kernel $K(\boldsymbol{x})$ are all true. In addition, we need the following assumptions to explore the large sample properties of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$.

($A1$) The bandwidth matrix $\boldsymbol{H}$ is symmetric, positive definite. As $n \to \infty$. $\boldsymbol{H} \to \mathbf{0}$. The ratio $\lambda_{\text{max}}(\boldsymbol{H})/\lambda_{\text{min}}(\boldsymbol{H})$ is bounded above, and $n|\boldsymbol{H}|\lambda^2_{\text{min}}(\boldsymbol{H}) \to \infty$ as $n \to \infty$. Here $\lambda_{\text{max}}(\boldsymbol{H})$ and $\lambda_{\text{min}}(\boldsymbol{H})$ are the maximum eigenvalue and the minimum eigenvalue of $\boldsymbol{H}$, respectively.

($A2$) $K(\boldsymbol{x})$ is Lipschitz continuous. That is, there exists $L > 0$, such that

$$|K(\boldsymbol{x}_1) - K(\boldsymbol{x}_2)| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \quad \forall \boldsymbol{x}_1,\ \boldsymbol{x}_2.$$

($A3$) $n \int \rho_n(\boldsymbol{x})d\boldsymbol{x} \to \rho_I$ as $n \to \infty$.

($A4$) There exists a constant C, such that $n \int |\rho_n(\boldsymbol{x})|d\boldsymbol{x} \leq C$.

($A5$) For any sequence $\varepsilon_n > 0$ satisfying $n^{1/d}\varepsilon_n \to \infty$,

$$n \int_{\|\boldsymbol{x}\| \geq \varepsilon_n} |\rho_n(\boldsymbol{x})|d\boldsymbol{x} \to 0 \quad as \quad n \to \infty.$$

In assumption ($A1$), $\boldsymbol{H} \to \mathbf{0}$ means that every entry of $\boldsymbol{H}$ goes to 0. Since $\boldsymbol{H}$ is symmetric and positive definite, $\boldsymbol{H} \to \mathbf{0}$ is equivalent to $\lambda_{\text{max}}(\boldsymbol{H}) \to 0$. Also, $n|\boldsymbol{H}|\lambda^2_{\text{min}}(\boldsymbol{H}) \to \infty$ is equivalent to $\frac{1}{n|\boldsymbol{H}|}\boldsymbol{H}^{-2} \to \mathbf{0}$, because the maximum eigenvalue of $\frac{1}{n|\boldsymbol{H}|}\boldsymbol{H}^{-2}$ is $\frac{1}{n|\boldsymbol{H}|\lambda^2_{\text{min}}(\boldsymbol{H})}$. Given a point $\boldsymbol{x}$, the "local bandwidth region" corresponding to $\boldsymbol{H}$ is $\{\boldsymbol{X} \in \mathbb{R}^2 : K\left(\boldsymbol{H}^{-1}(\boldsymbol{X} - \boldsymbol{x})\right) > 0\}$, i.e.,

$$\{\boldsymbol{X} \in \mathbb{R}^2 : (\boldsymbol{X} - \boldsymbol{x})^T \boldsymbol{H}^{-2}(\boldsymbol{X} - \boldsymbol{x}) < 1\}.$$

This is an ellipse centered at $\boldsymbol{x}$, with $\lambda_{\text{max}}(\boldsymbol{H})$ as a half of the length of its longest axis, and $\lambda_{\text{min}}(\boldsymbol{H})$ as a half of the length of its shortest axis. Since $\lambda_{\text{max}}(\boldsymbol{H})/\lambda_{\text{min}}(\boldsymbol{H})$ is bounded above, the ellipse can not degenerate to a flat shape as $n \to \infty$. In addition, all the eigenvalues of $\boldsymbol{H}$ go to 0 at the same rate. This implies that $|\boldsymbol{H}|$ is a quantity of order $O(\lambda^d_{\text{max}}(\boldsymbol{H}))$ because $|\boldsymbol{H}|$ is equal to the product of all the eigenvalues of $\boldsymbol{H}$. The condition $n|\boldsymbol{H}|\lambda^2_{\text{min}}(\boldsymbol{H}) \to \infty$ requires that every eigenvalue of $\boldsymbol{H}$ should converge to

0 at a rate slower than $O(1/n^{1/(d+2)})$. For the univariate case, $(A1)$ reduces to $\boldsymbol{H} \to 0$ and $n\boldsymbol{H}^3 \to \infty$ as $n \to \infty$.

Assumption $(A4)$ implies that the integral of $\int |\rho_n(\boldsymbol{x})|d\boldsymbol{x}$ should vanish as $n \to \infty$, and the vanishing speed should not be slower than $O(1/n)$. Assumption $(A4)$ and assumption $(A5)$ further imply that the integral of $|\rho_n(\boldsymbol{x})|$ is essentially dominated by the values of $\rho_n(\boldsymbol{x})$ near the origin $\mathbf{0}$. The quantity ρ_I may be called the total correlation. $(A3)$ and $(A5)$ imply that for any fixed $\varepsilon > 0$,

$$\rho_I = \lim_{n\to\infty} n \int_{||\boldsymbol{x}||<\varepsilon} \rho_n(\boldsymbol{x})d\boldsymbol{x}.$$

The reason is

$$\begin{aligned} \lim_{n\to\infty} n \int_{||\boldsymbol{x}||<\varepsilon} \rho_n(\boldsymbol{x})d\boldsymbol{x} &= \lim_{n\to\infty} n \int \rho_n(\boldsymbol{x})d\boldsymbol{x} - \lim_{n\to\infty} n \int_{||\boldsymbol{x}||\geq\varepsilon} \rho_n(\boldsymbol{x})d\boldsymbol{x} \\ &= \rho_I \quad \text{(by } (A3) \text{ and } (A5)). \end{aligned}$$

Two examples of valid correlation functions that satisfy $(A3)$ to $(A5)$ are

$$\rho_n(\boldsymbol{x}) = \exp(-\alpha n^{1/d}||\boldsymbol{x}||),$$

and

$$\rho_n(\boldsymbol{x}) = \frac{1}{1 + \alpha(n^{1/d}||\boldsymbol{x}||)^2},$$

with $\alpha > 0$ in both cases.

In general, if $\rho_n(\boldsymbol{x}) = \rho(n^{1/d}\boldsymbol{x})$ and $\rho(\boldsymbol{x})$ is a fixed valid correlation function, which is continuous everywhere except at a finite number of points and absolutely integrable in $\mathbb{R}^d$ (i.e., $\int |\rho(\boldsymbol{x})|d\boldsymbol{x} < \infty$), then it is easy to check that $\rho_n(\boldsymbol{x})$ satisfies assumptions $(A3)$ to $(A5)$.

2.2 Two useful lemmas

Before we continue our discussion, we need to introduce some additional notation. We use $\mathbf{1}_d$ to denote a $(d \times 1)$ vector with every entry equal to 1, and $\mathbf{1}_{d\times d}$ to denote a $(d \times d)$ matrix with every entry equal to 1.

The theoretical derivations in this chapter depend on the two important lemmas presented in this section.

The following lemma is basically a collection of some equations in Ruppert and Wand (1994), but with some differences in notation and in the conditions for $\boldsymbol{H}$.

Lemma 2.1 *Under assumption (A1),*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x}) &= f(\boldsymbol{x})+o_p(1) && (2.3)\\
\frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x})(\boldsymbol{X}_i-\boldsymbol{x}) &= \mu_2(K)\boldsymbol{H}^2\nabla_f(\boldsymbol{x})+\boldsymbol{H}^2o_p(\mathbf{1}_d) && (2.4)\\
\frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x})(\boldsymbol{X}_i-\boldsymbol{x})(\mathbf{X}_i-\boldsymbol{x})^T &= \mu_2(K)f(\boldsymbol{x})\boldsymbol{H}^2+\boldsymbol{H}o_p(\mathbf{1}_{d\times d})\boldsymbol{H} && (2.5)\\
\frac{1}{n}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x}) &= |\boldsymbol{H}|^{-1}f(\boldsymbol{x})\mu(K^2)+o_p(|\boldsymbol{H}|^{-1}) && (2.6)\\
\frac{1}{n}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x})(\boldsymbol{X}_i-\boldsymbol{x}) &= |\boldsymbol{H}|^{-1}o_p(\mathbf{1}_d) && (2.7)\\
\frac{1}{n}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x})(\boldsymbol{X}_i-\boldsymbol{x})(\mathbf{X}_i-\boldsymbol{x})^T &= |\boldsymbol{H}|^{-1}o_p(\mathbf{1}_{d\times d}) && (2.8)
\end{aligned}$$

Proof: We only prove (2.4). The proofs for the remaining expressions are analogous. Let

$$\boldsymbol{W}_n=\frac{1}{n}\boldsymbol{H}^{-2}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\mathbf{X}_{\mathrm{i}}-\boldsymbol{x})(\boldsymbol{X}_{\mathrm{i}}-\boldsymbol{x}).$$

Then (2.4) is equivalent to

$$\boldsymbol{W}_n=\mu_2(K)\nabla_f(\boldsymbol{x})+o_p(\mathbf{1}_d).$$

It is sufficient to show $\mathrm{E}(\boldsymbol{W}_n)\to\mu_2(K)\nabla_f(\boldsymbol{x})$ and $\mathrm{Var}(\boldsymbol{W}_n)\to\mathbf{0}$ as $n\to\infty$. The expectation

$$\begin{aligned}
\mathrm{E}(\boldsymbol{W}_n) &= \boldsymbol{H}^{-2}\mathrm{E}\{K_{\boldsymbol{H}}(\boldsymbol{X}-\boldsymbol{x})(\boldsymbol{X}-\boldsymbol{x})\}\\
&= \boldsymbol{H}^{-2}\int|\boldsymbol{H}|^{-1}K(\boldsymbol{H}^{-1}(\boldsymbol{u}-\boldsymbol{x}))(\boldsymbol{u}-\boldsymbol{x})f(\boldsymbol{u})d\boldsymbol{u}.
\end{aligned}$$

By taking the transformation

$$\boldsymbol{u}=\boldsymbol{x}+\boldsymbol{H}\boldsymbol{v}$$

and using a Taylor series expansion, we have

$$\begin{aligned}
\mathrm{E}(\boldsymbol{W}_n) &= \boldsymbol{H}^{-1}\int K(\boldsymbol{v})\boldsymbol{v}f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{v})d\boldsymbol{v} \\
&= \boldsymbol{H}^{-1}\int K(\boldsymbol{v})\boldsymbol{v}\left\{f(\boldsymbol{x})+\boldsymbol{v}^T\boldsymbol{H}\nabla_f(\boldsymbol{x})+\frac{1}{2}\boldsymbol{v}^T\boldsymbol{H}\mathcal{H}_f(\boldsymbol{x}+\theta\boldsymbol{H}\boldsymbol{v})\boldsymbol{H}\boldsymbol{v}\right\}d\boldsymbol{v},
\end{aligned}$$

for some $0 \le \theta \le 1$. Then,

$$\mathrm{E}(\boldsymbol{W}_n) = \mu_2(K)\nabla_f(\boldsymbol{x}) + \frac{1}{2}\boldsymbol{H}^{-1}\int K(\boldsymbol{v})\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{H}\mathcal{H}_f(\boldsymbol{x}+\theta\boldsymbol{H}\boldsymbol{v})\boldsymbol{H}\boldsymbol{v}d\boldsymbol{v}.$$

Since $f(\boldsymbol{x})$ has been assumed to have second order continuous partial derivatives in Ω, which is a bounded closed set, every entry of $\mathcal{H}_f(\boldsymbol{x})$ must be bounded in Ω. So, $\mathcal{H}_f(\boldsymbol{x})$ must be bounded in Ω. Let $\|\mathcal{H}_f(\boldsymbol{x}+\theta\boldsymbol{H}\boldsymbol{v})\| < C$. Then, the absolute values of the components of the second term of $\mathrm{E}(\boldsymbol{W}_n)$ are bounded by the corresponding components of

$$\frac{1}{2}K_{\max}CV_0\|\boldsymbol{H}^{-1}\|\|\boldsymbol{H}\|^2\mathbf{1}_d = \frac{1}{2}K_{\max}CV_0\frac{\lambda_{\max}(\boldsymbol{H})}{\lambda_{\min}(\boldsymbol{H})}\lambda_{\max}(\boldsymbol{H})\mathbf{1}_d,$$

where $K_{\max}$ is the maximum value of the kernel, and V_0 is the volume of the unit ball in $I\!R^d$. By assumption $(A1)$, $\frac{\lambda_{\max}(\boldsymbol{H})}{\lambda_{\min}(\boldsymbol{H})}$ is also bounded, and $\lambda_{\max}(\boldsymbol{H}) \to 0$ (because $\boldsymbol{H} \to \mathbf{0}$). Therefore, the second term of $\mathrm{E}(\boldsymbol{W}_n)$ converges to $\mathbf{0}$. So

$$\mathrm{E}(\boldsymbol{W}_n) \to \mu_2(K)\nabla_f(\boldsymbol{x}).$$

as $n \to \infty$.

Next, we need to prove that $\mathrm{Var}(\boldsymbol{W}_n)$ converges to 0 as $n \to \infty$. Notice that

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{W}_n) &= \frac{1}{n}\boldsymbol{H}^{-2}\mathrm{Var}\{K_{\boldsymbol{H}}(\boldsymbol{X}-\boldsymbol{x})(\boldsymbol{X}-\boldsymbol{x})\}\boldsymbol{H}^{-2} \\
&\le \frac{1}{n}\boldsymbol{H}^{-2}\mathrm{E}\{K^2_{\boldsymbol{H}}(\boldsymbol{X}-\boldsymbol{x})(\boldsymbol{X}-\boldsymbol{x})(\boldsymbol{X}-\boldsymbol{x})^T\}\boldsymbol{H}^{-2} \\
&= \frac{1}{n}|\boldsymbol{H}|^{-2}\boldsymbol{H}^{-2}\int K^2(\boldsymbol{H}^{-1}(\boldsymbol{u}-\boldsymbol{x}))(\boldsymbol{u}-\boldsymbol{x})(\boldsymbol{u}-\boldsymbol{x})^Tf(\boldsymbol{u})d\boldsymbol{u}\boldsymbol{H}^{-2} \\
&= \frac{1}{n}|\boldsymbol{H}|^{-1}\boldsymbol{H}^{-1}\int K^2(\boldsymbol{v})\boldsymbol{v}\boldsymbol{v}^Tf(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{v})d\boldsymbol{v}\boldsymbol{H}^{-1} \\
&= \frac{1}{n}|\boldsymbol{H}|^{-1}\boldsymbol{H}^{-1}\int K^2(\boldsymbol{v})\boldsymbol{v}\boldsymbol{v}^T\{f(\boldsymbol{x})+o(1)\}d\boldsymbol{v}\boldsymbol{H}^{-1} \\
&= \frac{1}{n}|\boldsymbol{H}|^{-1}\boldsymbol{H}^{-2}\mu_2(K^2)\{f(\boldsymbol{x})+o(1)\}
\end{aligned}$$

The condition $\frac{1}{n}|\boldsymbol{H}|^{-1}\boldsymbol{H}^{-2} \to 0$, which is equivalent to $n|\boldsymbol{H}|\lambda^2_{\min}(\boldsymbol{H}) \to \infty$ in assumption (A1), is sufficient to force $\text{Var}(\boldsymbol{W}_n) \to 0$. So under assumption (A1), $\text{Var}(\boldsymbol{W}_n) \to 0$. This completes the proof of (2.4).

■

In Ruppert and Wand (1994), the bandwidth matrix is denoted as $\boldsymbol{H}^{1/2}$, while it is denoted as $\boldsymbol{H}$ here. Our assumption (A1) is more restrictive than the corresponding assumption given by Ruppert and Wand (1994, (A3) on page 1349). Not many details of proof are given in their paper and we have been unable to reproduce their results under their assumptions. Consequently, we have provided a proof of (2.4) to illustrate why we need to impose a stronger assumption.

The next lemma is new. It is crucial for deriving formulas for the asymptotic mean squared errors of kernel smoothing estimators when data are correlated.

Lemma 2.2 *Let*

$$\begin{aligned}
s_1(\boldsymbol{x}, n) &\equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j),\\
s_2(\boldsymbol{x}, n) &\equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j)(\boldsymbol{X}_i - \boldsymbol{x}),\\
s_3(\boldsymbol{x}, n) &\equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j)(\boldsymbol{X}_i - \boldsymbol{x})(\boldsymbol{X}_j - \boldsymbol{x})^T.
\end{aligned}$$

Then, for any interior point $\boldsymbol{x}$ in the design region Ω, under assumptions (A1) to (A5),

$$\begin{aligned}
s_1(\boldsymbol{x}, n) &= \frac{f(\boldsymbol{x})\mu(K^2)(1 + f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}), && (2.9)\\
s_2(\boldsymbol{x}, n) &= \frac{1}{n|\boldsymbol{H}|}o_p(\mathbf{1}_d), && (2.10)\\
s_3(\boldsymbol{x}, n) &= \frac{1}{n|\boldsymbol{H}|}o_p(\mathbf{1}_{d\times d}). && (2.11)
\end{aligned}$$

To prove this lemma, we need to establish four propositions in advance.

Proposition 2.1 *Under assumptions (A1) and (A5), for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} n|\boldsymbol{H}| \int_{\|\boldsymbol{u}\|\geq\varepsilon} |\rho_n(\boldsymbol{H}\boldsymbol{u})|d\boldsymbol{u} = 0. \qquad (2.12)$$

Proof:

$$n|\boldsymbol{H}|\int_{\|\boldsymbol{u}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{u})|d\boldsymbol{u} = n\int_{\|\boldsymbol{H}^{-1}\boldsymbol{v}\|\geq\varepsilon}|\rho_n(\boldsymbol{v})|d\boldsymbol{v}.$$

Note that $\| \boldsymbol{H}^{-1}\boldsymbol{v} \| \geq \varepsilon$ implies that $\| \boldsymbol{H}^{-1} \| \| \boldsymbol{v} \| \geq \varepsilon$, i.e.,

$$\| \boldsymbol{v} \| \geq \varepsilon / \| \boldsymbol{H}^{-1} \| = \lambda_{\min}(\boldsymbol{H})\varepsilon.$$

So,

$$n|\boldsymbol{H}|\int_{\|\boldsymbol{u}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{u})|d\boldsymbol{u} \leq n\int_{\|\boldsymbol{v}\|\geq\lambda_{\min}(\boldsymbol{H})\varepsilon}|\rho_n(\boldsymbol{v})|d\boldsymbol{v}.$$

By assumption $(A1)$, $n|\boldsymbol{H}| \to \infty$ and $\lambda_{\max}(\boldsymbol{H})/\lambda_{\min}(\boldsymbol{H})$ is bounded above. This implies $n^{1/d}\lambda_{\min}(\boldsymbol{H})\varepsilon \to \infty$. Hence the proposition follows from $(A5)$.

■

Proposition 2.2 *Under assumptions* $(A4)$, *there exists a constant* C_1, *such that*

$$n^2|\boldsymbol{H}|^2\int\int\int_{\|\boldsymbol{u}\|\leq 1,\,\|\boldsymbol{v}\|\leq 1,\,\|\boldsymbol{w}\|\leq 1}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{w}))|d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w} \leq C_1. \quad (2.13)$$

Proof:

$$\begin{aligned}
& n^2|\boldsymbol{H}|^2\int\int\int_{\|\boldsymbol{u}\|\leq 1,\,\|\boldsymbol{v}\|\leq 1,\,\|\boldsymbol{w}\|\leq 1}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{w}))|d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w} \\
= & \int_{\|\boldsymbol{u}\|\leq 1}\{n|\boldsymbol{H}|\int_{\|\boldsymbol{v}\|\leq 1}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))|d\boldsymbol{v}\}\{n|\boldsymbol{H}|\int_{\|\boldsymbol{w}\|\leq 1}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{w}))|d\boldsymbol{w}\}d\boldsymbol{u} \\
\leq & \int_{\|\boldsymbol{u}\|\leq 1}\{n\int|\rho_n(\boldsymbol{v})|d\boldsymbol{v}\}\{n\int|\rho_n(\boldsymbol{w})|d\boldsymbol{w}\}d\boldsymbol{u} \\
\leq & \int_{\|\boldsymbol{u}\|\leq 1}C^2d\boldsymbol{u} \quad \text{(by assumption } (A4)) \\
= & C^2 \times (\text{Volume of the ball } \{\boldsymbol{u} : \| \boldsymbol{u} \| \leq 1\}) \\
\equiv & C_1.
\end{aligned}$$

■

Proposition 2.3 *Under the assumptions* $(A1)$ *to* $(A5)$,

$$\lim_{n\to\infty} n|\boldsymbol{H}|\int\int K(\boldsymbol{u})K(\boldsymbol{v})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u}d\boldsymbol{v} = \mu(K^2)\rho_I. \quad (2.14)$$

Proof: Let

$$g_n(\boldsymbol{v}) = n|\boldsymbol{H}| \int K(\boldsymbol{u})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u}.$$

Then,

$$n|\boldsymbol{H}| \int\int K(\boldsymbol{u})K(\boldsymbol{v})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u}d\boldsymbol{v} = \int K(\boldsymbol{v})g_n(\boldsymbol{v})d\boldsymbol{v}. \tag{2.15}$$

Notice that

$$\begin{aligned} |g_n(\boldsymbol{v})| &\leq K_{\max}\{n|\boldsymbol{H}| \int |\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))|d\boldsymbol{u}\} \\ &\leq K_{\max}\{n \int |\rho_n(\boldsymbol{t})|d\boldsymbol{t}\}, \end{aligned}$$

where $K_{\max}$ is the maximum value of the kernel. By assumption $(A4)$, $n\int|\rho_n(\boldsymbol{t})|d\boldsymbol{t}$ is bounded. Hence $g_n(\boldsymbol{v})$ is also bounded. Because of the continuity of $K(\cdot)$ and $\rho_n(\cdot)$, $K(\boldsymbol{u})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))$ is continuous in $\{(\boldsymbol{u},\boldsymbol{v}) :\parallel \boldsymbol{u} \parallel\leq 1, \parallel \boldsymbol{v} \parallel\leq 1\}$. Therefore $g_n(\boldsymbol{v})$ is continuous in $\{\boldsymbol{v} :\parallel \boldsymbol{v} \parallel\leq 1\}$.

Next, we prove

$$\lim_{n\to\infty} g_n(\boldsymbol{v}) = K(\boldsymbol{v})\rho_I \quad \forall \;\; \boldsymbol{v} \in \{\boldsymbol{v} :\parallel \boldsymbol{v} \parallel< 1\}. \tag{2.16}$$

Because $\{\boldsymbol{u} :\parallel \boldsymbol{u} \parallel< 1\}$ is an open set, for any fixed $\boldsymbol{v}$ satisfying $\parallel \boldsymbol{v} \parallel< 1$, we can always choose a small enough $\varepsilon > 0$, such that $\{\boldsymbol{u} :\parallel \boldsymbol{u}-\boldsymbol{v} \parallel< \varepsilon\}$ is contained in $\{\boldsymbol{u} :\parallel \boldsymbol{u} \parallel< 1\}$. Hence

$$\begin{aligned} g_n(\boldsymbol{v}) &= n|\boldsymbol{H}|K(\boldsymbol{v}) \int_{\|\boldsymbol{u}-\boldsymbol{v}\|<\varepsilon} \rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u} \\ &\quad + n|\boldsymbol{H}| \int_{\|\boldsymbol{u}-\boldsymbol{v}\|<\varepsilon} (K(\boldsymbol{u}) - K(\boldsymbol{v}))\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u} \\ &\quad + n|\boldsymbol{H}| \int_{\|\boldsymbol{u}-\boldsymbol{v}\|\geq\varepsilon} K(\boldsymbol{u})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u} \\ &\equiv I_1 + I_2 + I_3. \end{aligned} \tag{2.17}$$

The first term

$$\begin{aligned} I_1 &= K(\boldsymbol{v})\{n \int_{\|\boldsymbol{H}^{-1}\boldsymbol{t}\|\leq\varepsilon} \rho_n(\boldsymbol{t})d\boldsymbol{t}\} \\ &= K(\boldsymbol{v})\{n \int \rho_n(\boldsymbol{t})d\boldsymbol{t} - n \int_{\|\boldsymbol{H}^{-1}\boldsymbol{t}\|\geq\varepsilon} \rho_n(\boldsymbol{t})d\boldsymbol{t}\}. \end{aligned}$$

Notice that

$$\begin{aligned}|n\int_{\|\boldsymbol{H}^{-1}\boldsymbol{t}\|\geq\varepsilon}\rho_n(\boldsymbol{t})d\boldsymbol{t}| &\leq n|\boldsymbol{H}|\int_{\|\boldsymbol{u}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{u})|d\boldsymbol{u}\\ &\to 0,\ as\ n\to\infty\ \text{(by Propososition 2.1).}\end{aligned}$$

By assumption $(A3)$,

$$\lim_{n\to\infty} n\int\rho_n(\boldsymbol{t})d\boldsymbol{t}=\rho_I.$$

So we have

$$\lim_{n\to\infty} I_1 = K(\boldsymbol{v})\rho_I. \tag{2.18}$$

Now consider the second term. Since $K(\boldsymbol{x})$ is a Lipschitz continuous,

$$\begin{aligned}|I_2| &\leq L\varepsilon\{n|\boldsymbol{H}|\int_{\|\boldsymbol{u}-\boldsymbol{v}\|\leq\varepsilon}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))|d\boldsymbol{u}\}\\ &= L\varepsilon\{n\int_{\|\boldsymbol{H}^{-1}\boldsymbol{t}\|<\varepsilon}|\rho_n(\boldsymbol{t})|d\boldsymbol{t}\}\\ &\leq LC\varepsilon\ \text{(by assumption (A4)).}\end{aligned} \tag{2.19}$$

Then consider the third term,

$$\begin{aligned}|I_3| &\leq K_M\{n|\boldsymbol{H}|\int_{\|\boldsymbol{u}-\boldsymbol{v}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))|d\boldsymbol{u}\}\\ &= K_M\{n|\boldsymbol{H}|\int_{\|\boldsymbol{u}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}(\boldsymbol{u}))|d\boldsymbol{u}\}\\ &\to 0\ as\ n\to\infty\ \text{(by Proposition 2.1).}\end{aligned} \tag{2.20}$$

Since ε can be arbitrarily small, (2.16) is true.

So far, we have shown that $g_n(\boldsymbol{v})$ is bounded, continuous, and converges to $K(\boldsymbol{v})\rho_I$ everywhere in $\{\boldsymbol{v}:\|\ \boldsymbol{v}\ \|<1\}$. Due to the continuity of $K(\boldsymbol{v})$, $K(\boldsymbol{v})g_n(\boldsymbol{v})$ is also bounded, continuous, and converges to $K^2(\boldsymbol{v})\rho_I$ everywhere in $\{\boldsymbol{v}:\|\ \boldsymbol{v}\ \|\leq 1\}$, except on the boundary of the ball. Also notice that the set $\{\boldsymbol{v}:\|\ \boldsymbol{v}\ \|\leq 1\}$ has a finite Lebesgue measure. Hence in the following step, we can apply the Lebesgue bounded convergence theorem to change the order of the limit and the integral.

$$\lim_{n\to\infty}\int_{\|\boldsymbol{v}\|\leq 1} K(\boldsymbol{v})g_n(\boldsymbol{v})d\boldsymbol{v}=\int_{\|\boldsymbol{v}\|\leq 1} K(\boldsymbol{v})\lim_{n\to\infty} g_n(\boldsymbol{v})d\boldsymbol{v}=\mu(K^2)\rho_I.$$

Hence the proof is completed .

■

Proposition 2.4 *Let*

$$B_n(\boldsymbol{x}) \equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} K_{\boldsymbol{H}}(\mathbf{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j), \tag{2.21}$$

$$C_n(\boldsymbol{x}) \equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} |K_{\boldsymbol{H}}(\mathbf{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j)|. \tag{2.22}$$

Then, for any interior point $\boldsymbol{x}$ in the design region Ω, under assumptions $(A1)$ to $(A5)$,

$$B_n(\boldsymbol{x}) = \frac{f^2(\boldsymbol{x})\mu(K^2)\rho_I}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}), \tag{2.23}$$

$$C_n(\boldsymbol{x}) = O_p(\frac{1}{n|\boldsymbol{H}|}). \tag{2.24}$$

Proof: Let

$$\psi_{ij}(\boldsymbol{x}) \equiv K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j).$$

Then

$$B_n(\boldsymbol{x}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\psi_{ij}(\boldsymbol{x}),$$

and

$$C_n(\boldsymbol{x}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}|\psi_{ij}(\boldsymbol{x})|.$$

Next, we will prove (2.23), which is equivalent to

$$n|\boldsymbol{H}|B_n(\boldsymbol{x}) = f^2(\boldsymbol{x})\mu(K^2)\rho_I + o_p(1). \tag{2.25}$$

We need to show

$$\lim_{n\to\infty} \mathrm{E}(n|\boldsymbol{H}|B_n(\boldsymbol{x})) = f^2(\boldsymbol{x})\mu(K^2)\rho_I,$$

and

$$\lim_{n\to\infty} \mathrm{Var}(n|\boldsymbol{H}|B_n(\boldsymbol{x})) = 0.$$

Notice that

$$
\begin{aligned}
& \mathrm{E}(n|\boldsymbol{H}|B_n(\boldsymbol{x})) \\
= \;& n^{-1}|\boldsymbol{H}|(n^2-n)\mathrm{E}(K_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{x})\rho_n(\boldsymbol{X}_i-\boldsymbol{X}_j)) \\
= \;& (1-n^{-1})n|\boldsymbol{H}|\int\int K_{\boldsymbol{H}}(\boldsymbol{\zeta}-\boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{\eta}-\boldsymbol{x})\rho_n(\boldsymbol{\zeta}-\boldsymbol{\eta})f(\boldsymbol{\zeta})f(\boldsymbol{\eta})d\boldsymbol{\zeta}d\boldsymbol{\eta} \\
= \;& (1-n^{-1})n|\boldsymbol{H}|\int\int K(\boldsymbol{u})K(\boldsymbol{v})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{u}) \\
& \quad \cdot f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} \\
= \;& f^2(\boldsymbol{x})\{n|\boldsymbol{H}|\int\int K(\boldsymbol{u})K(\boldsymbol{v})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u}d\boldsymbol{v}\}(1+o(1)) \\
\rightarrow \;& f^2(\boldsymbol{x})\mu(K^2)\rho_I \quad \textit{as} \;\; n\rightarrow\infty \quad \text{(by Proposition 2.3).}
\end{aligned}
$$

And we have

$$
\mathrm{Var}(n|\boldsymbol{H}|B_n(\boldsymbol{x})) = 4n^{-2}|\boldsymbol{H}|^2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{l=1}^{n-1}\sum_{m=l+1}^{n}\mathrm{Cov}(\psi_{ij}(\boldsymbol{x}),\psi_{lm}(\boldsymbol{x})).
$$

Now consider the value of $\mathrm{Cov}(\psi_{ij}(\boldsymbol{x}),\psi_{lm}(\boldsymbol{x}))$ according to the following three complete and exclusive cases.

Case 1: i, j, l, m are all distinct.

$\mathrm{Cov}(\psi_{ij}(\boldsymbol{x}),\psi_{lm}(\boldsymbol{x}))=0$, because $\psi_{ij}(\boldsymbol{x})$ and $\psi_{lm}(\boldsymbol{x})$ are independent.

Case 2: One of the i and j is equal to one of the l and m.

The total number of such terms $< n^3$. Notice that $K(\cdot)$ and $\rho_n(\cdot)$ are symmetric, hence all terms in this case have the same value. Without loss of generality, assume $i=l$ and $j\neq m$. Denote $K_M\equiv\max(K(\boldsymbol{x}))$ and $f_M\equiv\max_{\boldsymbol{x}\in\Omega}f(\boldsymbol{x})$.

$$
\begin{aligned}
& \mathrm{Cov}(\psi_{ij},\psi_{im}) \\
= \;& \mathrm{E}(\psi_{ij}\psi_{im})-\mathrm{E}(\psi_{ij})\mathrm{E}(\psi_{im}) \\
= \;& \mathrm{E}(\psi_{ij}\psi_{im})-(\mathrm{E}(\psi_{ij}))^2 \\
\leq \;& \mathrm{E}(\psi_{ij}\psi_{im}) \\
= \;& \int\int\int K_{\boldsymbol{H}}^2(\boldsymbol{\zeta}-\boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{\eta}-\boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{\tau}-\boldsymbol{x})\rho_n(\boldsymbol{\zeta}-\boldsymbol{\eta})\rho_n(\boldsymbol{\zeta}-\boldsymbol{\tau})f(\boldsymbol{\zeta})f(\boldsymbol{\eta})
\end{aligned}
$$

$$
\begin{aligned}
&\cdot f(\boldsymbol{\tau})d\boldsymbol{\zeta}d\boldsymbol{\eta}d\boldsymbol{\tau} \\
= \ & |\boldsymbol{H}|^{-1}\int\int\int K^2(\boldsymbol{u})K(\boldsymbol{v})K(\boldsymbol{w})\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{w}))f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{u}) \\
&\cdot f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{v})f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{w})d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w} \\
\leq \ & n^{-2}|\boldsymbol{H}|^{-3}K_M^4 f_M^3\{n^2|\boldsymbol{H}|^2\int\int\int_{||\boldsymbol{u}||\leq 1,\, ||\boldsymbol{v}||\leq 1,\, ||\boldsymbol{w}||\leq 1}|\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))\rho_n(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{w}))| \\
&\cdot d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w}\} \\
\leq \ & \frac{C_1K_M^4f_M^3}{n^2|\boldsymbol{H}|^3} \quad \text{(by Proposition 2.2)} \\
\equiv \ & \frac{\alpha_1}{n^2|\boldsymbol{H}|^3}.
\end{aligned}
$$

Case 3: $i = l < j = m$.

The total number of such terms is $(n^2-n)/2$.

$$
\begin{aligned}
&\mathrm{Cov}(\psi_{ij},\psi_{lm}) \\
= \ & \mathrm{Var}(K_{\boldsymbol{H}}(\boldsymbol{X}_i-\boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{x})\rho_n(\boldsymbol{X}_i-\boldsymbol{X}_j)) \\
\leq \ & \int\int K_{\boldsymbol{H}}^2(\boldsymbol{\zeta}-\boldsymbol{x})K_{\boldsymbol{H}}^2(\boldsymbol{\eta}-\boldsymbol{x})\rho_n^2(\boldsymbol{\zeta}-\boldsymbol{\eta})f(\boldsymbol{\zeta})f(\boldsymbol{\eta})d\boldsymbol{\zeta}d\boldsymbol{\eta} \\
\leq \ & |\boldsymbol{H}|^{-2}\int\int K^2(\boldsymbol{u})K^2(\boldsymbol{v})\rho_n^2(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{u})f(\boldsymbol{x}+\boldsymbol{H}\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} \\
\leq \ & n^{-1}|\boldsymbol{H}|^{-3}K_M^4f_M^2\int_{||\boldsymbol{v}||\leq 1}\{n|\boldsymbol{H}|\int_{||\boldsymbol{u}||\leq 1}\rho_n^2(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u}\}d\boldsymbol{v}.
\end{aligned}
$$

Since

$$
n|\boldsymbol{H}|\int_{||\boldsymbol{u}||\leq 1}\rho_n^2(\boldsymbol{H}(\boldsymbol{u}-\boldsymbol{v}))d\boldsymbol{u} \leq n\int|\rho_n(\boldsymbol{t})|d\boldsymbol{t} \leq C,
$$

we get

$$
\mathrm{Cov}(\psi_{ij},\psi_{im}) \leq \frac{CK_M^4f_M^2}{n|\boldsymbol{H}|^3} \times (\text{Volume of the ball } \{\boldsymbol{u} : ||\ \boldsymbol{u}\ ||\leq 1\}) \equiv \frac{\alpha_2}{n|\boldsymbol{H}|^3}.
$$

Using the covariance results of all the three cases, we have

$$
\begin{aligned}
\mathrm{Var}(n|\boldsymbol{H}|B_n(\boldsymbol{x})) \ &\leq \ \frac{4|\boldsymbol{H}|^2}{n^2}(n^3\cdot\frac{\alpha_1}{n^2|\boldsymbol{H}|^3}+\frac{n^2-n}{2}\cdot\frac{\alpha_2}{n|\boldsymbol{H}|^3}) \\
&= \ \frac{4}{n|\boldsymbol{H}|}(\alpha_1+\frac{1-n^{-1}}{2}\alpha_2) \\
&\to \ 0, \quad \text{as } n\to\infty.
\end{aligned}
$$

Hence we have proved (2.23).

By very similar arguments, we can show (2.24). The details are omitted .

■

Next, we are ready to prove Lemma 2.2. Notice that the sum

$$s_1(\boldsymbol{x},n) = \frac{1}{n^2}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) + B_n(\boldsymbol{x}).$$

From (2.6) of Lemma 2.1,

$$\frac{1}{n^2}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) = \frac{f(\boldsymbol{x})\mu(K^2)}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}).$$

From (2.23) of Proposition 2.4,

$$B_n(\boldsymbol{x}) = \frac{f^2(\boldsymbol{x})\mu(K^2)\rho_I}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}).$$

Hence (2.9) holds.

Equation (2.10) is equivalent to

$$\boldsymbol{c}^T \boldsymbol{s}_2(\boldsymbol{x},n) = o_p(\frac{1}{n|\boldsymbol{H}|}),$$

for any fixed vector $\boldsymbol{c} \in I\!R^d$. Notice that

$$\begin{aligned} \boldsymbol{c}^T \boldsymbol{s}_2(\boldsymbol{x},n) &= \frac{1}{n^2}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})\boldsymbol{c}^T(\boldsymbol{X}_i - \boldsymbol{x})1_{\|\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\| \le 1} + \\ &\quad \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j \ne i} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{x})\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j)\boldsymbol{c}^T(\boldsymbol{X}_i - \boldsymbol{x}) \\ &\quad \cdot 1_{\{\|\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\| \le 1\}}, \end{aligned}$$

where $1_{\{\|\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x})\| \le 1\}}$ is an indicator function, equal to 1 if $\| \boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x}) \| \le 1$ and equal to 0 otherwise. Thus,

$$|\boldsymbol{c}^T \boldsymbol{s}_2(\boldsymbol{x},n)| \le \| \boldsymbol{c} \| \| \boldsymbol{H} \| \{\frac{1}{n^2}\sum_{i=1}^{n} K^2_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) + C_n(\boldsymbol{x})\}.$$

By (2.6) of Lemma 2.1 and (2.24) of Proposition 2.4,

$$|\boldsymbol{c}^T \boldsymbol{s}_2(\boldsymbol{x},n)| = O_p(\frac{\| \boldsymbol{H} \|}{n|\boldsymbol{H}|}) = o_p(\frac{1}{n|\boldsymbol{H}|}).$$

Hence (2.10) holds.

Realizing that (2.11) is equivalent to

$$c_1^T s_3(\boldsymbol{x}, n) c_2 = o_p(\frac{1}{n|\boldsymbol{H}|}),$$

for any fixed vectors c_1, $c_2 \in I\!\!R^d$, the proof of (2.11) is then similar to the proof of (2.10) and thus omitted .

■

2.3 Asymptotic MSE of kernel smoothing estimators when data are correlated

In this section, we will derive the asymptotic Mean Squared Errors (MSE) of three different kernel smoothing estimators for model (2.1).

2.3.1 Asymptotic MSE of the local linear estimator

The local linear estimator of the mean function at $\boldsymbol{x}$ is written as

$$\hat{m}(\boldsymbol{x}, \boldsymbol{H}) = e_1^T (\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}})^{-1} \boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{Y}, \tag{2.26}$$

where

$$\boldsymbol{X}_{\boldsymbol{x}} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{x})^T \\ \vdots & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{x})^T \end{pmatrix},$$

$$\boldsymbol{W}_{\boldsymbol{x}} = \mathrm{diag}(K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{x}), K_{\boldsymbol{H}}(\boldsymbol{X}_2 - \boldsymbol{x}), \ldots, K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{x})),$$

and

$$e_1 = (1, 0, \ldots, 0)^T \in I\!\!R^{d+1}.$$

The conditional mean of the estimator is

$$\mathrm{E}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = e_1^T (\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}})^{-1} \boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{m},$$

with

$$\boldsymbol{m} = (m(\boldsymbol{X}_1), \ldots, m(\boldsymbol{X}_n))^T.$$

The conditional variance of the estimator is

$$\begin{aligned} &\mathrm{Var}(\hat{m}(\boldsymbol{x}, \boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \\ = \ & e_1^T (\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}})^{-1} \boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{V} \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}} (\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}})^{-1} e_1, \end{aligned} \tag{2.27}$$

where $\boldsymbol{V}$ is the variance matrix of the vector of random errors.

We have the following theorem.

Theorem 2.1 *Let $\boldsymbol{x}$ be a fixed interior point in the support of $f(\boldsymbol{x})$, the density function of $\boldsymbol{X}$. Let $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ be the local linear estimator of $m(\boldsymbol{x})$. Under assumptions $(A1)$ to $(A5)$,*

$$E\{\hat{m}(\boldsymbol{x}, \boldsymbol{H}) - m(\boldsymbol{x}) | \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\} = \frac{1}{2}\mu_2(K) \ \mathrm{tr}(\boldsymbol{H}^2 \mathcal{H}_m(\boldsymbol{x})) + o_p(\mathrm{tr}(\boldsymbol{H}^2)), \tag{2.28}$$

where $\mathcal{H}_m(\boldsymbol{x})$ is the Hessian matrix of $m(\boldsymbol{x})$, and

$$\mathrm{Var}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = \frac{\sigma^2 \mu(K^2)(1 + f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}| f(\boldsymbol{x})} + o_p(\frac{1}{n|\boldsymbol{H}|}). \tag{2.29}$$

Proof: First, the bias of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is not related to the correlation structure of random errors. Hence the bias result for correlated data is the same as the bias result for uncorrelated data, which was given by Ruppert and Wand (1994).

Based on Ruppert and Wand (1994, page 1353),

$$(\frac{1}{n}\boldsymbol{X}_{\boldsymbol{x}}^T \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}})^{-1} = \begin{pmatrix} f^{-1}(\boldsymbol{x}) + o_p(1) & -f^{-2}(\boldsymbol{x}) \nabla_f^T(\boldsymbol{x}) + o_p(\boldsymbol{1}^T) \\ -f^{-2}(\boldsymbol{x}) \nabla_f(\boldsymbol{x}) + o_p(\boldsymbol{1}) & \{\mu_2(K) f(\boldsymbol{x}) \boldsymbol{H}^2\}^{-1} + \boldsymbol{H} o_p(\boldsymbol{1}_{d\times d}) \boldsymbol{H} \end{pmatrix}. \tag{2.30}$$

The proof of the above equation needs the result of Lemma 2.1. Now with (2.27) in mind, we need to find an asymptotic expression for $\frac{1}{n^2}\boldsymbol{X}_{\boldsymbol{x}} \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{V} \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}}$. Notice that

$$\frac{1}{n^2}\boldsymbol{X}_{\boldsymbol{x}} \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{V} \boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}} = \sigma^2 \begin{pmatrix} s_1(\boldsymbol{x}, n) & s_2^T(\boldsymbol{x}, n) \\ s_2(\boldsymbol{x}, n) & s_3(\boldsymbol{x}, n) \end{pmatrix},$$

where $s_1(\boldsymbol{x}, n)$, $s_2(\boldsymbol{x}, n)$, and $s_3(\boldsymbol{x}, n)$ are defined in Lemma 2.2. Hence by Lemma 2.2,

$$\frac{1}{n^2}\boldsymbol{X_x W_x V W_x X_x} = \frac{\sigma^2}{n|\boldsymbol{H}|}\begin{pmatrix} f(\boldsymbol{x})\mu(K^2)(1+f(\boldsymbol{x})\rho_I)+o_p(1) & o_p(\boldsymbol{1}^T) \\ o_p(\boldsymbol{1}) & o_p(\boldsymbol{1}_{d\times d}) \end{pmatrix}. \quad (2.31)$$

From (2.30) and (2.31),

$$\begin{aligned} & \operatorname{Var}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H})|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\} \\ = \; & \boldsymbol{e}_1^T(n^{-1}\boldsymbol{X}_{\boldsymbol{x}}^T\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{X}_{\boldsymbol{x}})^{-1}(n^{-2}\boldsymbol{X}_{\boldsymbol{x}}^T\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{V}\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{X}_{\boldsymbol{x}})(n^{-1}\boldsymbol{X}_{\boldsymbol{x}}^T\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{X}_{\boldsymbol{x}})^{-1}\boldsymbol{e}_1 \\ = \; & \frac{\sigma^2\mu(K^2)(1+f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|f(\boldsymbol{x})} + o_p(\frac{1}{n|\boldsymbol{H}|}). \end{aligned}$$

This completes the proof.

■

When data are uncorrelated, $\rho_I = 0$, and equation (2.29) reduces to the asymptotic variance expression of the local linear estimator given by Ruppert and Wand (1994).

According to assumption $(A1)$, $\frac{\lambda_{\max}(\boldsymbol{H})}{\lambda_{\min}(\boldsymbol{H})}$ is bounded above. This implies that all the eigenvalues of $\boldsymbol{H}$ have the same convergence rate as $\lambda_{\max}(\boldsymbol{H})$ when $n \to \infty$. So $|\boldsymbol{H}|$, which is equal to the product of all the eigenvalues of $\boldsymbol{H}$, has the same convergence rate as $O(\lambda_{\max}^d(\boldsymbol{H}))$. We get

$$O_p\left(\frac{1}{n|\boldsymbol{H}|}\right) = O_p\left(\frac{1}{n\lambda_{\max}^d(\boldsymbol{H})}\right). \quad (2.32)$$

Since

$$d\lambda_{\min}^2(\boldsymbol{H}) = d\lambda_{\min}(\boldsymbol{H}^2) \le \operatorname{tr}(\boldsymbol{H}^2) \le d\lambda_{\max}(\boldsymbol{H}^2) = d\lambda_{\max}^2(\boldsymbol{H}),$$

$\operatorname{tr}(\boldsymbol{H}^2)$ and $\lambda_{\max}^2(\boldsymbol{H})$ have the same convergence rate when $n \to \infty$. Thus,

$$O_p(\operatorname{tr}(\boldsymbol{H}^2)) = O_p(\lambda_{\max}^2(\boldsymbol{H})). \quad (2.33)$$

From Theorem 2.1, as $n \to \infty$, the asymptotic MSE of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ has a convergence rate of $O_p\left(\lambda_{\max}^4(\boldsymbol{H}) + \frac{1}{n\lambda_{\max}^d(\boldsymbol{H})}\right)$.

To ensure $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ to be a consistent estimator of $m(\cdot)$ at the point $\boldsymbol{x}$, $\lambda_{\max}^4(\boldsymbol{H}) + \frac{1}{n\lambda_{\max}^d(\boldsymbol{H})}$ must converge to 0. So $\lambda_{\max}(\boldsymbol{H})$ must converge to 0 in an order slower than

$O(1/n^{1/d})$. As discussed in section 2.1, assumption ($A1$) requires that $\lambda_{\max}(\boldsymbol{H})$ converges to 0 in a rate slower than $O(1/n^{1/(d+2)})$. Because $O(1/n^{1/(d+2)})$ is slower than $O(1/n^{1/d})$, $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is a consistent estimator of $m(\boldsymbol{x})$ if $\lambda_{\max}(\boldsymbol{H})$ converges to 0 in a rate slower than $O(1/n^{1/(d+2)})$. To get the optimal rate at which AMSE($\hat{m}(\boldsymbol{x}, \boldsymbol{H})$) goes to 0, we can take

$$\lambda_{\max}(\boldsymbol{H}) = O(\frac{1}{n^{1/(d+4)}}).$$

So the optimal rate at which AMSE($\hat{m}(\boldsymbol{x}, \boldsymbol{H})$) goes to 0 is $O_p(\frac{1}{n^{4/(d+4)}})$.

To understand the effect of the bandwidth matrix $\boldsymbol{H}$ on $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$, let $\boldsymbol{H} = h\boldsymbol{H}_0$, with $\boldsymbol{H}_0$ being a fixed matrix, not changing with the sample size. Geometrically, $\boldsymbol{H}_0$ controls the shape and the orientation of the elliptical bandwidth region corresponding to $\boldsymbol{H}$, while h controls the size of this region. From Theorem 2.1, the leading term of the asymptotic bias of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is

$$\frac{1}{2}\mu_2(K)h^2 \operatorname{tr}(\boldsymbol{H}_0^2\mathcal{H}_m(\boldsymbol{x})).$$

So the bias is small when h is small. In other words, the bias is small when the elliptical bandwidth region is small. The bias is also related to the Hessian matrix $\mathcal{H}_m(\boldsymbol{x})$, whose entries are the second order derivatives of the mean function at $\boldsymbol{x}$, measuring the curvatures of the mean function at $\boldsymbol{x}$ along various directions. Roughly speaking, when the curvatures of the mean function are small, the bias is also small.

From Theorem 2.1, the leading term of the asymptotic (conditional) variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is

$$\frac{\sigma^2\mu(K^2)}{n|\boldsymbol{H}|}(1/f(\boldsymbol{x}) + \rho_I).$$

When σ^2, the variance of the errors, is small, the variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is also small. When the sample size n is large, the variance is small. Notice that $|\boldsymbol{H}|$ is proportional to the Lebesgue measure of the elliptical bandwidth region corresponding to $\boldsymbol{H}$. So when the size of the bandwidth region is large, $|\boldsymbol{H}|$ is large, so that the variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$

is small. The design density at $\boldsymbol{x}$ also affects the variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$. When $f(\boldsymbol{x})$ is larger (the number of design points near x is larger), the variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is smaller. The integral term ρ_I may be called the total correlation of the errors. Compared to the case of uncorrelated errors, a positive correlation of the errors ($\rho_I > 0$) causes a larger variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$. The larger the positive correlation, the larger the variance of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$.

In summary, Theorem 2.1 identifies the factors that affect the asymptotic mean squared error of $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$. In section 2.4, Theorem 2.1 will be used to derive formulas of asymptotic optimal bandwidth for the local linear estimator.

2.3.2 Asymptotic MSE of the Priestley-Chao estimator

For the Priestley-Chao estimator, Altman (1990) gave the asymptotic mean squared error expression in the univariate case for fixed, equally spaced designs. Here, we consider the multivariate case under a uniformly random design.

The Priestley-Chao estimator of $m(\boldsymbol{x})$ is

$$\hat{m}(\boldsymbol{x}, \boldsymbol{H}) = \frac{b}{n} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) Y_i,$$

where b is the geometric volume of the bounded design region Ω.

Theorem 2.2 *Suppose that the design is uniformly random. Let $\boldsymbol{x}$ be a fixed interior point in Ω. Let $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ be the Priestley-Chao estimator of $m(\boldsymbol{x})$. Under assumptions $(A1)$ to $(A5)$,*

$$\mathrm{E}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H}) - m(\boldsymbol{x}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = \frac{1}{2}\mu_2(K)\ \mathrm{tr}(\boldsymbol{H}^2 \mathcal{H}_m(\boldsymbol{x})) + o_p(\mathrm{tr}(\boldsymbol{H}^2)), \qquad (2.34)$$

where $\mathcal{H}_m(\boldsymbol{x})$ is the Hessian matrix of $m(\boldsymbol{x})$, and

$$\mathrm{Var}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = \frac{\sigma^2 \mu(K^2)(1 + \rho_I / b)}{n|\boldsymbol{H}|/b} + o_p(\frac{1}{n|\boldsymbol{H}|}). \qquad (2.35)$$

Proof: The proof is sketched as follows. The conditional expectation of the estimator $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is

$$\mathrm{E}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = \frac{b}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})m(\boldsymbol{X}_i).$$

Note that for the uniformly random design, the design density $f(\boldsymbol{x}) = 1/b$.

$$\begin{aligned} & \mathrm{E}\left(\frac{b}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})m(\boldsymbol{X}_i)\right) \\ = \; & b\int K_{\boldsymbol{H}}(\boldsymbol{u} - \boldsymbol{x})m(\boldsymbol{u})f(\boldsymbol{u})\boldsymbol{u} \\ = \; & \int K(\boldsymbol{v})m(\boldsymbol{x} + \boldsymbol{H}\boldsymbol{v})d\boldsymbol{v} \\ = \; & \int K(\boldsymbol{v})\{m(\boldsymbol{x}) + \nabla_m(\boldsymbol{x})'\boldsymbol{H}\boldsymbol{v} + \frac{1}{2}\boldsymbol{v}'\boldsymbol{H}\mathcal{H}_m(\boldsymbol{x})\boldsymbol{H}\boldsymbol{v} + \boldsymbol{v}'\boldsymbol{H}o(\mathbf{1}_{d\times d})\boldsymbol{H}\boldsymbol{v}\}d\boldsymbol{v} \\ = \; & m(\boldsymbol{x}) + \frac{1}{2}\mu_2(K)\ \mathrm{tr}(\mathcal{H}_m(\boldsymbol{x})\boldsymbol{H}^2) + o(\mathrm{tr}(\boldsymbol{H}^2)). \end{aligned}$$

It can be shown further that Var $\left(\frac{b}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})m(\boldsymbol{X}_i)\right) = o(\mathrm{tr}^2(\boldsymbol{H}^2))$. Hence the equation (2.34) holds.

Now consider the conditional expectation of the estimator $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$. Note that

$$\mathrm{Var}\{\hat{m}(\boldsymbol{x}, \boldsymbol{H})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = b^2\sigma^2 s_1(\boldsymbol{x}, n),$$

where $s_1(\boldsymbol{x}, n)$ is defined in Lemma 2.2. Then, equation (2.35) follows immediately from (2.9) of Lemma 2.2 .

■

Next, we try to compare our result in the uniformly random design with the result in the fixed equally spaced design given by Altman (1990). The purpose for doing this is to study possible similarities and differences between a random design and a fixed design through this simple case.

Consider the univariate case with the fixed design in the region $\Omega = [0, 1]$ $(b = 1)$. Following Altman (1990), the design points are $x_i = \frac{i}{n}$, and the correlation between the error at $x_i = \frac{i}{n}$ and the error at $x_j = \frac{j}{n}$ is

$$\rho_n(|x_i - x_j|) = \rho(|i - j|).$$

with ρ as a fixed correlation function. This means that the correlation of the errors is a function of the difference between the indices of the corresponding design points. The above equation is equivalent to

$$\rho_n(|x_i - x_j|) = \rho(n|x_i - x_j|). \tag{2.36}$$

Now for the uniformly random design case, we introduce a spatial correlation function, $\rho_n(t) = \rho(nt)$, which denotes the correlation of the random errors at two locations with distance t in between when the sample size is n. As a result, the correlation structure in (2.36) for the fixed design is naturally extended to the random design case. According to Theorem 2.1, under a uniformly random design, the asymptotic mean squared error at a design point x is

$$\text{AMSE}(x, H) = \left\{\frac{1}{2}\mu_2(K)H^2\mathcal{H}_m(x)\right\}^2 + \frac{\sigma^2\mu(K^2)(1+\rho_I)}{nH}. \tag{2.37}$$

For the fixed equally spaced design, the asymptotic mean squared error at a design point x, given by Altman (1990), is

$$\text{AMSE}(x, H) = \left\{\frac{1}{2}\mu_2(K)H^2\mathcal{H}_m(x)\right\}^2 + \frac{\sigma^2\mu(K^2)(1+2R)}{nH}, \tag{2.38}$$

with $R = \sum_{k=1}^{\infty}\rho(k)$. So no difference can be found in the first terms, or the asymptotic bias terms for the two different designs. But there does exist a difference in the second terms, or the asymptotic variance terms. Note that $2R = \sum_{k=-\infty}^{\infty}\rho(k) - \rho(0)$. So in the fixed, equally spaced design case, correlations between random errors affect the variance term via the sum of correlations at all possible nonzero lags. In the uniformly random design case, however, the correlations between random errors affect the variance term via ρ_I, which is approximately equal to the integral of the correlation function over the entire space multiplied by n (for large n). In the following, we will establish an inequality between R and ρ_I for certain types of correlation functions.

Let $\rho(|t|)$ be a decreasing nonnegative function of $|t|$. In spatial statistics, this assumption is often appropriate. It means that two datum points that are closer are more strongly positively correlated. Under this assumption, we have

$$n\int_{-1}^{1}\rho(nt)dt = 2n\sum_{k=1}^{n}\int_{\frac{k-1}{n}}^{\frac{k}{n}}\rho(nt)dt > 2n\sum_{k=1}^{n}\rho(n\cdot\frac{k}{n})\frac{1}{n} = 2\sum_{k=1}^{n}\rho(k). \tag{2.39}$$

On the other hand,

$$\begin{aligned} n\int_{-1}^{1}\rho(nt)dt &= 2n\sum_{k=1}^{n}\int_{\frac{k-1}{n}}^{\frac{k}{n}}\rho(nt)dt < 2n\sum_{k=1}^{n}\rho(n\cdot\frac{k-1}{n})\frac{1}{n} \\ &= 2\sum_{k=1}^{n}\rho(k-1) \leq 2+2\sum_{k=1}^{n-1}\rho(k). \end{aligned}$$

So

$$2+2\sum_{k=1}^{n-1}\rho(k) > n\int_{-1}^{1}\rho(nt)dt > 2\sum_{k=1}^{n}\rho(k).$$

Let $n \to \infty$, we obtain

$$2+2R \geq \rho_I \geq 2R. \tag{2.40}$$

Then, using the fact $\rho_I \geq 2R$, it can be easily seen from (2.37) and (2.38) that the asymptotic variance in a uniformly random design is not smaller than the asymptotic variance in a fixed, equally spaced design. Notice also that both designs yield the same asymptotic bias. So, for purpose of estimating the mean function, if a design needs to be chosen between the fixed, equally spaced design and the uniformly random design, one would prefer the fixed, equally spaced design, because it leads to a smaller asymptotic mean squared error for the estimator of the mean function point-wisely. We believe that the similarities and differences found between these two designs are also true for multivariate cases, and this is an interesting problem for further study.

As an example, let ρ be an exponentially decreasing function: $\rho(i) = \alpha^i$, with $0 < \alpha < 1$. Then

$$\rho_I = \lim_{n\to\infty} n\int_{-1}^{1}\rho(nt)dt = \lim_{n\to\infty} 2n\int_{0}^{1}\alpha^{nt}dt = -\frac{2}{\log(\alpha)},$$

and

$$2R = 2\sum_{i=1}^{\infty}\alpha^i = \frac{2\alpha}{1-\alpha}.$$

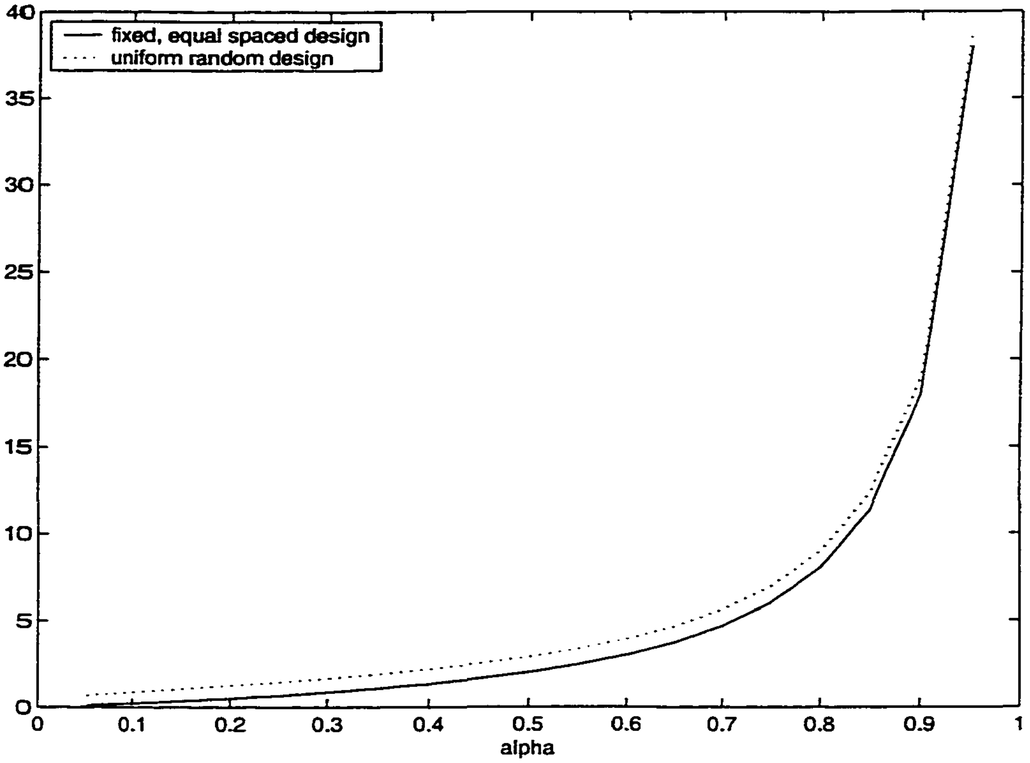

Figure 2.1 ρ_I and $2R$ versus α

Then it can be shown that for any $\alpha \in (0, 1)$, $\rho_I > 2R$, and

$$\lim_{\alpha \to 1^-} (\rho_I - 2R) = \lim_{\alpha \to 1^-} \left(-\frac{2}{\log(\alpha)} - \frac{2\alpha}{1-\alpha} \right) = 1.$$

In Figure 2.1, we plot ρ_I and $2R$ as a function of α. The dotted line represents ρ_I, and the solid line represents $2R$.

2.3.3 Asymptotic MSE of the Nadaraya-Watson estimator

The Nadaraya-Watson estimator is given by

$$\hat{m}(\boldsymbol{x}, \boldsymbol{H}) = \frac{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})Y_i}{\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})}. \tag{2.41}$$

To our knowledge, the asymptotic mean squared error of this estimator has not been studied when data are correlated. Unlike the Priestley-Chao estimator, the Nadaraya-

Watson estimator is asymptotically unbiased for any random design. The following theorem addresses the asymptotic MSE of this estimator.

Theorem 2.3 *Let $\boldsymbol{x}$ be a fixed interior point in the support of $f(\boldsymbol{x})$, the density function of $\boldsymbol{X}$. Let $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ represent the Nadaraya-Watson estimator of $m(\boldsymbol{x})$. Under assumptions $(A1)$ to $(A5)$,*

$$\begin{aligned} & \mathrm{E}\{\hat{m}(\boldsymbol{x},\boldsymbol{H}) - m(\boldsymbol{x}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \\ = \; & \frac{1}{2}\mu_2(K)\ \mathrm{tr}(\boldsymbol{H}^2 \mathcal{H}_m(\boldsymbol{x})) + \mu_2(K)/f(\boldsymbol{x}) \nabla_m^T(\boldsymbol{x}) \boldsymbol{H}^2 \nabla_f(\boldsymbol{x}) + o_p(\mathrm{tr}(\boldsymbol{H}^2)), \quad (2.42) \end{aligned}$$

where $\mathcal{H}_m(\boldsymbol{x})$ is the Hessian matrix of $m(\boldsymbol{x})$, and

$$\mathrm{Var}\{\hat{m}(\boldsymbol{x},\boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = \frac{\sigma^2 \mu(K^2)(1 + f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}| f(\boldsymbol{x})} + o_p(\frac{1}{n|\boldsymbol{H}|}). \quad (2.43)$$

Proof: Since the asymptotic conditional bias, given by (2.42), is not related to the correlations between random errors, the bias result for correlated data is the same as the bias result for uncorrelated data, which was given by Härdle and Müller (2000). We only need to show (2.43). Note that

$$\mathrm{Var}\{\hat{m}(\boldsymbol{x},\boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} = \frac{s_1(\boldsymbol{x}, n)}{\left(\frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})\right)^2},$$

where $s_1(\boldsymbol{x}, n)$ is defined in Lemma 2.2. From (2.9) of Lemma 2.2,

$$s_1(\boldsymbol{x}, n) = \frac{f(\boldsymbol{x})\mu(K^2)(1 + f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}).$$

From (2.3) of Lemma 2.1,

$$\frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) = f(\boldsymbol{x}) + o_p(1),$$

which implies

$$\frac{1}{\left(\frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})\right)^2} = \frac{1}{f^2(\boldsymbol{x})} + o_p(1).$$

Hence (2.43) is true.

■

From Theorem 2.1 and Theorem 2.3, we see that the local linear estimator and the Nadaraya-Watson estimator have the same asymptotic variance, but their asymptotic biases are different. The Nadaraya-Watson estimator, due to the second term of its bias $\mu_2(K)/f(\boldsymbol{x})\nabla_m^T(\boldsymbol{x})\boldsymbol{H}^2\nabla_f(\boldsymbol{x})$, is inferior to the local linear estimator. Even if the mean function is linear, the Nadaraya-Watson estimator still has a second order bias. The Nadaraya-Watson estimator is not design-adaptive, a term used by Fan (1992), in the sense that its bias depends on the design density.

2.4 Asymptotic optimal bandwidth

In a multivariate case with uncorrelated data, the formula for the asymptotic local optimal bandwidth for a local linear estimator can be found in Fan and Gijbels (1992a). In this section, for kernel smoothing estimators in case of correlated data, we will consider the asymptotic optimal bandwidth, both locally and globally. The following proposition is needed for further discussion.

Proposition 2.5 *Let* $\mathcal{S}_0 = \{\boldsymbol{X} : \boldsymbol{X} \text{ is a } d \times d \text{ matrix}, \ |\boldsymbol{X}| > 0\}$. *Assuming that* $\boldsymbol{A}$ *is a* $d \times d$ *symmetric matrix (either positive definite or negative definite), and* c_1, c_2 *are positive numbers, then the solution to the optimization problem,*

$$\min_{\boldsymbol{X} \in \mathcal{S}_0} L(\boldsymbol{X}) = \left\{ c_1 \operatorname{tr}^2(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}) + \frac{c_2}{|\boldsymbol{X}|} \right\}, \tag{2.44}$$

is the matrix $\boldsymbol{X}$ *satisfying*

$$\boldsymbol{X}\boldsymbol{X}^T = \left\{ \frac{c_2 |\tilde{\boldsymbol{A}}|^{1/2}}{4 c_1 d} \right\}^{2/(d+4)} (\tilde{\boldsymbol{A}})^{-1}, \tag{2.45}$$

with

$$\tilde{\boldsymbol{A}} = \begin{cases} \boldsymbol{A} & \text{if } \boldsymbol{A} \text{ is positive definite,} \\ -\boldsymbol{A} & \text{if } \boldsymbol{A} \text{ is negative definite.} \end{cases} \tag{2.46}$$

Proof: For $\boldsymbol{X} \in \mathcal{S}_0$, $0 < L(\boldsymbol{X}) < \infty$. If we take $\boldsymbol{X} = \delta \boldsymbol{I}_d$, then $\lim_{\delta \to 0^+} L(\boldsymbol{X}) = +\infty$. So the problem given by (2.44) is well defined, while the problem of maximizing the objective function $L(\boldsymbol{X})$ in $\mathcal{S}_0$ is not.

Assuming that $g(\cdot)$ is a scalar function of a matrix $\boldsymbol{X} = (x_{ij})_{d \times d}$, we define the partial derivative of g with respect to $\boldsymbol{X}$ as

$$\frac{\partial g}{\partial \boldsymbol{X}} = \left(\frac{\partial g}{\partial x_{ij}} \right)_{d \times d}.$$

Note that

$$\frac{\partial |\boldsymbol{X}|}{\partial \boldsymbol{X}} = |\boldsymbol{X}|(\boldsymbol{X}^T)^{-1},$$

and

$$\frac{\partial \mathrm{tr}^2(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X})}{\partial \boldsymbol{X}} = 4 \, \mathrm{tr}(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}) \boldsymbol{A} \boldsymbol{X}.$$

Then the solution to the minimization problem (2.44) satisfies

$$4c_1 \, \mathrm{tr}(\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}) \boldsymbol{A} \boldsymbol{X} - \frac{c_2}{|\boldsymbol{X}|} (\boldsymbol{X}^T)^{-1} = 0.$$

That is

$$\boldsymbol{X}\boldsymbol{X}^T = \frac{c_2}{4c_1 \, \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T)|\boldsymbol{X}|} \boldsymbol{A}^{-1}, \tag{2.47}$$

i.e.,

$$\boldsymbol{X}\boldsymbol{X}^T = \alpha \boldsymbol{A}^{-1}, \tag{2.48}$$

with

$$\alpha = \frac{c_2}{4c_1 \, \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T)|\boldsymbol{X}|}.$$

Equation (2.47) has a solution if and only if $\boldsymbol{A}$ is either positive definite or negative definite.

In the proof, we consider the case when $\boldsymbol{A}$ is negative definite only. In this case, $\tilde{\boldsymbol{A}} = -\boldsymbol{A}$ is positive definite, and $\mathrm{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T)$ is negative. From (2.47), $|\boldsymbol{X}|$ has to be positive so that both sides of (2.47) are positive definite matrices. So $\alpha < 0$. Let $\tilde{\alpha} = -\alpha$. Then

$$|\boldsymbol{X}| = \sqrt{|\boldsymbol{X}\boldsymbol{X}^T|} = \sqrt{|\tilde{\alpha}(\tilde{\boldsymbol{A}})^{-1}|} = (\tilde{\alpha})^{d/2} |\tilde{\boldsymbol{A}}|^{-1/2}. \tag{2.49}$$

Then, substitute (2.48) and (2.49) into (2.47) to obtain

$$\tilde{\alpha} = \frac{c_2}{4c_1(d\tilde{\alpha})\tilde{\alpha}^{d/2}|\tilde{\boldsymbol{A}}|^{-1/2}}.$$

This gives

$$\tilde{\alpha} = \left\{\frac{c_2|\tilde{\boldsymbol{A}}|^{1/2}}{4c_1 d}\right\}^{2/(d+4)}.$$

So

$$\boldsymbol{X}\boldsymbol{X}^T = \tilde{\alpha}(\tilde{\boldsymbol{A}})^{-1} = \left\{\frac{c_2|\tilde{\boldsymbol{A}}|^{1/2}}{4c_1 d}\right\}^{2/(d+4)} (\tilde{\boldsymbol{A}})^{-1}.$$

Similar arguments can be applied to the case when $\boldsymbol{A}$ is positive definite.

■

The following proposition is more useful in this section, because it can be directly applied to bandwidth matrices which are symmetric, positive definite.

Proposition 2.6 *Let* $\mathcal{S} = \{\boldsymbol{X} : \boldsymbol{X}$ is a $d \times d$ symmetric, positive definite matrix$\}$. *Assuming that* $\boldsymbol{A}$ *is a* $d \times d$ *symmetric matrix (either positive definite or negative definite), and* c_1, c_2 *are positive numbers, then the solution to the optimization problem,*

$$\min_{\boldsymbol{X} \in \mathcal{S}} L(\boldsymbol{X}) = \left\{c_1 \operatorname{tr}^2(\boldsymbol{X}^2\boldsymbol{A}) + \frac{c_2}{|\boldsymbol{X}|}\right\}, \tag{2.50}$$

is

$$\boldsymbol{X} = \left\{\frac{c_2|\tilde{\boldsymbol{A}}|^{1/2}}{4c_1 d}\right\}^{1/(d+4)} (\tilde{\boldsymbol{A}})^{-1/2}, \tag{2.51}$$

with

$$\tilde{\boldsymbol{A}} = \begin{cases} \boldsymbol{A} & \text{if } \boldsymbol{A} \text{ is positive definite,} \\ -\boldsymbol{A} & \text{if } \boldsymbol{A} \text{ is negative definite.} \end{cases} \tag{2.52}$$

Proof: Note that (2.51) implies (2.45), and that the objective function of problem (2.50) is equal to the objective function of problem (2.44) in the set $\mathcal{S}$. So the $\boldsymbol{X}$ given by (2.51) is a solution to problem (2.44). That is, the $\boldsymbol{X}$ given by (2.51) is optimal in $\mathcal{S}_0$. Since $\mathcal{S} \subset \mathcal{S}_0$, the $\boldsymbol{X}$ given by (2.51) must be optimal in $\mathcal{S}$. Hence the proposition is true.

■

For a given interior point $\boldsymbol{x}$ in the design region, let

$$\tilde{\mathcal{H}}_m(\boldsymbol{x}) = \begin{cases} \mathcal{H}_m(\boldsymbol{x}) & \text{if } \mathcal{H}_m(\boldsymbol{x}) \text{ is positive definite,} \\ -\mathcal{H}_m(\boldsymbol{x}) & \text{if } \mathcal{H}_m(\boldsymbol{x}) \text{ is negative definite.} \end{cases} \tag{2.53}$$

We consider the local linear estimator at first. Theorem 2.1 gives the formula of the (conditional) asymptotic mean squared error (AMSE) for the local linear estimator:

$$\text{AMSE}(\boldsymbol{x}, \boldsymbol{H}) = \frac{1}{4}\mu_2^2(K)\ \text{tr}^2(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{x})) + \frac{\sigma^2\mu(K^2)(1+f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|f(\boldsymbol{x})}. \tag{2.54}$$

Note that the asymptotic local optimal bandwidth is the one that minimizes the value of AMSE$(\boldsymbol{x}, \boldsymbol{H})$. By applying Proposition 2.6, the local optimal bandwidth for the local linear estimator is then

$$\boldsymbol{H}^l_{\text{opt}} = \left\{ \frac{\sigma^2\mu(K^2)(1+f(\boldsymbol{x})\rho_I)|\tilde{\mathcal{H}}_m(\boldsymbol{x})|^{1/2}}{n\ d\ \mu_2^2(K)f(\boldsymbol{x})} \right\}^{1/(d+4)} \left(\tilde{\mathcal{H}}_m(\boldsymbol{x})\right)^{-1/2} \equiv h^* \left(\tilde{\mathcal{H}}_m(\boldsymbol{x})\right)^{-1/2}. \tag{2.55}$$

Here the matrix $\left(\tilde{\mathcal{H}}_m(\boldsymbol{x})\right)^{-1/2}$ determines the shape and the orientation of the local optimal bandwidth region, while h^* decides its size. At a given point $\boldsymbol{x}$, the optimal bandwidth region can be written as

$$(\boldsymbol{X} - \boldsymbol{x})^T\tilde{\mathcal{H}}_m(\boldsymbol{x})(\boldsymbol{X} - \boldsymbol{x}) < r(\boldsymbol{x}),$$

where $r(\boldsymbol{x}) > 0$ is a value depending on $\boldsymbol{x}$. So the optimal local neighborhood at $\boldsymbol{x}$ has the same shape and orientation as the contour of the mean function $m(\cdot)$ near $\boldsymbol{x}$. In other words, The shape and orientation of the local optimal neighborhood at a given point is completely determined by the curvatures of the mean function $m(\cdot)$ at that point. Equation (2.55) may be useful for selecting local optimal bandwidth at a given point. However, point-wise bandwidth selection is extremely time consuming. So global bandwidth (the same bandwidth for all design points) selection is often used instead. To get an asymptotic global optimal bandwidth, we need to select $\boldsymbol{H}$ to minimize

$\text{AMISE}(\boldsymbol{H}) = \int \text{AMSE}(\boldsymbol{x}, \boldsymbol{H})w(\boldsymbol{x})dx$, where $w(\boldsymbol{x}) > 0$ is a weight function selected by the users. A possible choice would be selecting $w(\boldsymbol{x})$ as the design density $f(\boldsymbol{x})$. So, for the local linear estimator, we need to select $\boldsymbol{H}$ to minimize

$$\text{AMISE}(\boldsymbol{H}) = \int \left(\frac{1}{4}\mu_2^2(K) \ \text{tr}^2(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{x})) + \frac{\sigma^2\mu(K^2)(1 + f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|f(\boldsymbol{x})} \right) w(\boldsymbol{x})d\boldsymbol{x}. \quad (2.56)$$

Unfortunately, no closed form solution for this optimization problem can be found. In the following, we will look at several simpler situations.

If $m(\boldsymbol{x})$ is quadratic, then for any $\boldsymbol{x} \in \Omega$, $\mathcal{H}_m(\boldsymbol{x}) \equiv \mathcal{H}_m$ (a constant matrix). We assume that $\mathcal{H}_m$ is either positive definite, or negative definite. In this case, $\text{AMISE}(\boldsymbol{H})$ reduces to

$$\text{AMISE}(\boldsymbol{H}) = \frac{1}{4}\mu_2^2(K) \ \text{tr}^2(\boldsymbol{H}^2\mathcal{H}_m) + \frac{\sigma^2\mu(K^2)}{n|\boldsymbol{H}|} \left(\int w(\boldsymbol{x})/f(\boldsymbol{x})d\boldsymbol{x} + \rho_I \right),$$

Applying Proposition 2.6, we get the asymptotic global optimal bandwidth

$$\boldsymbol{H}^g_{\text{opt}} = \left(\frac{\sigma^2\mu(K^2)\,(\int w(\boldsymbol{x})/f(\boldsymbol{x})d\boldsymbol{x} + \rho_I)\,|\tilde{\mathcal{H}}_m|^{1/2}}{n \ d \ \mu_2^2(K)} \right)^{1/(d+4)} \tilde{\mathcal{H}}_m^{-1/2},$$

where $\tilde{\mathcal{H}}_m$ is equal to $\mathcal{H}_m$ if $\mathcal{H}_m$ is positive definite, and equal to $-\mathcal{H}_m$ if $\mathcal{H}_m$ is negative definite. In practice, this formula may not be useful, because if we knew that the mean function is quadratic, parametric regression (the method of universal kriging) would be a better choice. But it is interesting to look at what this formula suggests. The formula indicates a plain and somewhat disappointing fact: the global optimal bandwidth is generally a full bandwidth matrix. Speaking geometrically, when selecting the elliptical bandwidth region, its size, shape, and orientation should all be considered. Also, the formula indicates that the optimal elliptical bandwidth region should match the shape and the orientation of the contour of $m(\boldsymbol{x})$, which is usually unknown in practice.

In general, searching for a full (global) bandwidth matrix is also time consuming. In a bivariate case, when $m(\boldsymbol{x})$ is known to be an additive model, the global optimal bandwidth matrix turns out to be diagonal. To see this, let

$$m(\boldsymbol{x}) = m_1(x_1) + m_2(x_2),$$

with $\boldsymbol{x} = (x_1, x_2)^T$. Then

$$\mathcal{H}_m(\boldsymbol{x}) = \text{diag}(m_1''(x_1), m_2''(x_2)).$$

Denote $\boldsymbol{H}^2$ as

$$\boldsymbol{H}^2 \equiv \boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix}.$$

The AMISE($\boldsymbol{H}$) then simplifies to

$$\begin{aligned} &\text{AMISE}(\boldsymbol{H}) \\ &= \frac{1}{4}\mu_2^2(K)b_{11}^2 \int (m_1''(x_1))^2 w(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{4}\mu_2^2(K)b_{22}^2 \int (m_2''(x_2))^2 w(\boldsymbol{x})d\boldsymbol{x} \\ &\quad + \frac{1}{2}\mu_2^2(K)b_{11}b_{22} \int m_1''(x_1)m_2''(x_2)w(\boldsymbol{x})d\boldsymbol{x} + \frac{\sigma^2\mu(K^2)\left(\int w(\boldsymbol{x})/f(\boldsymbol{x})d\boldsymbol{x} + \rho_I\right)}{n\sqrt{b_{11}b_{22} - b_{12}^2}}. \end{aligned}$$

By setting its partial derivatives with respect to b_{11}, b_{22}, and b_{12} equal to 0, we get $b_{12} = 0$, and

$$\frac{b_{11}}{b_{22}} = \left(\frac{\int (m_2''(x_2))^2 w(\boldsymbol{x})d\boldsymbol{x}}{\int (m_1''(x_1))^2 w(\boldsymbol{x})d\boldsymbol{x}}\right)^{1/2}.$$

So the global optimal bandwidth $\boldsymbol{H}_{\text{opt}}^g$ is a diagonal matrix. Let $\boldsymbol{H}_{\text{opt}}^g = \text{diag}(h_1, h_2)$. Then,

$$\frac{h_1}{h_2} = \frac{\sqrt{b_{11}}}{\sqrt{b_{22}}} = \left(\frac{\int (m_2''(x_2))^2 w(\boldsymbol{x})d\boldsymbol{x}}{\int (m_1''(x_1))^2 w(\boldsymbol{x})d\boldsymbol{x}}\right)^{1/4}.$$

Here, $\int (m_2''(x_2))^2 w(\boldsymbol{x})d\boldsymbol{x}$ and $\int (m_1''(x_1))^2 w(\boldsymbol{x})d\boldsymbol{x}$ measure the average curvature of $m(\boldsymbol{x})$ with respect to x_1 and the average curvature of $m(\boldsymbol{x})$ with respect to x_2, respectively. The parameters h_1 and h_2 control the amount of smoothing along the directions of the x_1 axis and x_2 axis, separately. The above equation implies that if the curvature of $m(\boldsymbol{x})$ with respect to a certain variable is larger, then the amount of smoothing along the direction of the axis of that variable should be smaller.

In practice, searching for a diagonal global optimal bandwidth may be still computationally slow, because for a d-variate problem, d parameters in the bandwidth matrix need to be selected. To simplify the computation, we may restrict our attention to a special type bandwidth $\boldsymbol{H} = h\boldsymbol{I}_d$, which gives spherically shaped local bandwidth regions. This type of bandwidth matrix may be called a spherical bandwidth matrix. Searching for a spherical bandwidth is much faster, because only one parameter h needs to be selected. However, this kind of restricted search may lead to loss of efficiency for the estimator of $m(\boldsymbol{x})$. When we restrict attention to bandwidth matrices of this type, the minimizer of (2.56) can be explicitly solved. This gives the optimal global spherical bandwidth

$$h^g_{\text{opt}} = \left(\frac{\sigma^2 \; d \; \mu(K^2) \, (\int w(\boldsymbol{x})/f(\boldsymbol{x}) d\boldsymbol{x} + \rho_I)}{n \; \mu_2^2(K) \int \text{tr}^2(\mathcal{H}_m(\boldsymbol{x})) w(\boldsymbol{x}) d\boldsymbol{x}} \right)^{1/(d+4)} . \tag{2.57}$$

Applying formulas like (2.55) and (2.57) in bandwidth selection for real data requires prior estimates of several unknown quantities: the design density $f(\boldsymbol{x})$, the Hessian matrix of an unknown mean function $\mathcal{H}_m(\boldsymbol{x})$, and the variance of the errors and correlations between them. Plug-in type techniques may be useful here. The idea of the plug-in type approach is as follows. Start with a pilot (initial) fit of the mean function. Then use the result of the pilot fit to estimate the Hessian matrix of $m(\boldsymbol{x})$. Use residuals from this pilot fit to estimate the variance, and the correlations between the errors. Then apply those results to calculate the asymptotic optimal bandwidth. The major difficulty in using this approach is selecting a good pilot fit. This approach has been widely discussed in the case of uncorrelated data, but for the case of spatially correlated data, it remains unclear how it can be successfully applied. We will not pursue this approach in further depth.

Finally, we provide a brief discussion of the Priestley-Chao estimator and the Nadaraya-Watson estimator.

Under a uniformly random design, we consider the asymptotic local optimal bandwidth for the Priestley-Chao estimator. For a given interior point $\boldsymbol{x}$ in the design region, Theorem 2.2 gives the (conditional) asymptotic mean squared error (AMSE) formula for the Priestley-Chao estimator,

$$\text{AMSE}(\boldsymbol{x},\boldsymbol{H}) = \frac{1}{4}\mu_2^2(K)\ \text{tr}^2(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{x})) + \frac{\sigma^2\mu(K^2)(b+\rho_I)}{n|\boldsymbol{H}|}. \tag{2.58}$$

Applying Proposition 2.6, we get the asymptotic local optimal bandwidth

$$\boldsymbol{H}^l_{\text{opt}} = \left\{\frac{\sigma^2\mu(K^2)(b+\rho_I)|\tilde{\mathcal{H}}_m(\boldsymbol{x})|^{1/2}}{n\ d\ \mu_2^2(K)}\right\}^{1/(d+4)} \left(\tilde{\mathcal{H}}_m(\boldsymbol{x})\right)^{-1/2}, \tag{2.59}$$

where b is the Lebesgue measure of the design region Ω, and $\tilde{\mathcal{H}}_m(\boldsymbol{x})$ is defined in (2.53).

Then, under a general random design, we consider the asymptotic local optimal bandwidth for the Nadaraya-Watson estimator. For a given interior point $\boldsymbol{x}$ in the design region, by Theorem 2.3, we have

$$\begin{aligned}
&\text{AMSE}(\boldsymbol{x},\boldsymbol{H})\\
= &\left\{\frac{1}{2}\mu_2(K)\ \text{tr}(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{x})) + \frac{\mu_2(K)}{f(\boldsymbol{x})}\nabla_m^T(\boldsymbol{x})\boldsymbol{H}^2\nabla_f(\boldsymbol{x})\right\}^2 + \frac{\sigma^2\mu(K^2)(1+f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|f(\boldsymbol{x})}\\
= &\frac{1}{4}\mu_2^2(K)\ \text{tr}^2\left\{\boldsymbol{H}^2\left(\mathcal{H}_m(\boldsymbol{x}) + \frac{1}{f(\boldsymbol{x})}\left(\nabla_f(\boldsymbol{x})\nabla_m^T(\boldsymbol{x}) + \nabla_m(\boldsymbol{x})\nabla_f^T(\boldsymbol{x})\right)\right)\right\}\\
&+\frac{\sigma^2\mu(K^2)(1+f(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}|f(\boldsymbol{x})}.
\end{aligned}$$

Let

$$\mathcal{H}_{\text{NW}}(\boldsymbol{x}) \equiv \mathcal{H}_m(\boldsymbol{x}) + \frac{1}{f(\boldsymbol{x})}\left(\nabla_f(\boldsymbol{x})\nabla_m^T(\boldsymbol{x}) + \nabla_m(\boldsymbol{x})\nabla_f^T(\boldsymbol{x})\right),$$

and

$$\tilde{\mathcal{H}}_{\text{NW}}(\boldsymbol{x}) = \begin{cases} \mathcal{H}_{\text{NW}}(\boldsymbol{x}) & \text{if } \mathcal{H}_{\text{NW}}(\boldsymbol{x}) \text{ is positive definite.}\\ -\mathcal{H}_{\text{NW}}(\boldsymbol{x}) & \text{if } \mathcal{H}_{\text{NW}}(\boldsymbol{x}) \text{ is negative definite.}\end{cases} \tag{2.60}$$

By Proposition 2.6, the asymptotic local optimal bandwidth for the Nadaraya-Watson estimator is

$$\boldsymbol{H}^l_{\text{opt}} = \left\{\frac{\sigma^2\mu(K^2)(1+f(\boldsymbol{x})\rho_I)|\tilde{\mathcal{H}}_{\text{NW}}(\boldsymbol{x})|^{1/2}}{n\ d\ \mu_2^2(K)f(\boldsymbol{x})}\right\}^{1/(d+4)} \left(\tilde{\mathcal{H}}_{NW}(\boldsymbol{x})\right)^{-1/2}. \tag{2.61}$$

3 SELECTION OF BANDWIDTH WITH KNOWN COVARIANCES BETWEEN ERRORS

3.1 Introduction

In practice, kernel type smoothing techniques require the selection of the bandwidth $\boldsymbol{H}$. The bandwidth $\boldsymbol{H}$ controls the smoothness, bias and variance of the estimate of the mean function. When data are uncorrelated, various techniques have been developed for determining suitable values of the bandwidth from data. Among them, Mallows' C_L, cross-validation, and generalized cross-validation (Craven and Wahba, 1979) are popular choices. However, when data are correlated, the correlations between the errors have a disastrous effect on these methods. In the univariate case, when the design is fixed and equally spaced, Altman (1990) and Hart (1991) considered the Priestley-Chao estimator and the Gasser-Müller estimator respectively. They proposed new methods for bandwidth selection based on the estimates of the correlations between the errors, but it is not clear whether their bandwidth selection methods can be successfully extended to more general situations. In this chapter, we consider the more general case of multivariate predictors with random designs. We will consider the Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator, with emphasis on the last one because of its complexity and its nice asymptotic properties.

We use a simulated example to show that for correlated data the classical cross-validation method fails in selecting an appropriate bandwidth $\boldsymbol{H}$, and that modification of this type of methods is needed. We randomly generate 400 design points $\boldsymbol{X}_i =$

(X_{i1}, X_{i2}) in the rectangle: $0 < X_1 < 1, 0 < X_2 < 1$ according to a uniform distribution. The responses Y_i satisfy

$$Y_i = \sin(2\pi X_{i1}) + 4(X_{i2} - 0.5)^2 + \varepsilon_i,$$

where the random errors ε_i are normally distributed with variance $\sigma^2 = 0.25$ and covariance

$$\text{Cov}(\varepsilon_i, \varepsilon_j | \boldsymbol{X}_i, \boldsymbol{X}_j) = \sigma^2 \alpha^{20\sqrt{(\boldsymbol{X}_{i1} - \boldsymbol{X}_{j1})^2 + (\boldsymbol{X}_{i2} - \boldsymbol{X}_{j2})^2}},$$

with $\sigma = 0.5$ and $\alpha = 0.3$. Note that the variance of the mean function $m(\boldsymbol{X}) = \sin(2\pi X_1) + 4(X_2 - 0.5)^2$ is $\frac{53}{90}$. If we define the signal to noise ratio as the ratio of the variance of the mean function (the signal) to the variance of random errors (the noise), then it is approximately 2.36:1. Figure 3.1(a) displays the data $(\boldsymbol{X}_i, Y_i)$. Figure 3.1(b) displays the true mean function. We use the local linear regression with Epanechnikov kernel to uncover the true surface (the mean function). For simplicity, we consider 8 spherical bandwidth matrices $\boldsymbol{H} = h\boldsymbol{I}_2$, where h takes values from 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.45, and 0.5. The cross-validation (CV) criterion ignores the correlation among the errors. In this example, it selects $\boldsymbol{H} = 0.1\boldsymbol{I}_2$ and gives a very rough estimate of the surface (Figure 3.1 (c)). The corrected CV criterion (proposed in next section) selects $\boldsymbol{H} = 0.3\boldsymbol{I}_2$ and gives a quite smooth and more accurate estimate of the mean function (Figure 3.1 (d)). The corrected CV uses information about the correlation. In this example, we assume that the correlation matrix is completely known, although in practice, it needs to be estimated from the data.

In this chapter, we will address some of the issues raised by this example. We will investigate why the cross-validation type criteria that ignore the information of the correlations between random errors are not suitable for correlated data. We will modify the Mallows C_L, cross-validation, and generalized cross-validation so that they can be successfully applied to correlated data.

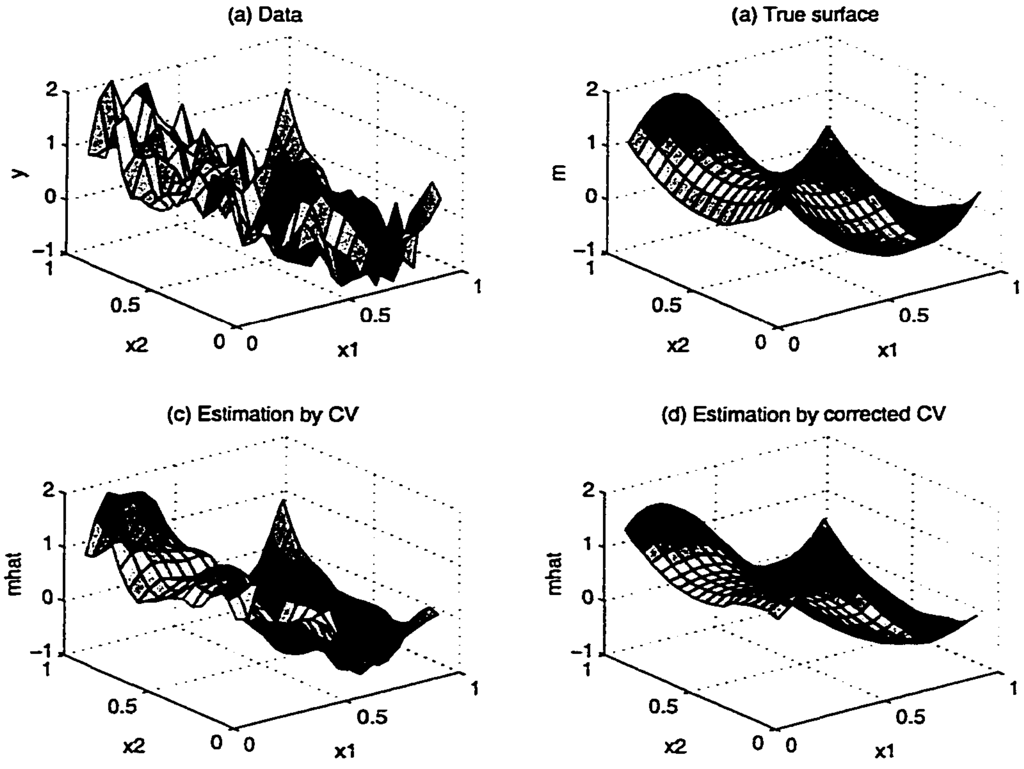

Figure 3.1 A simulated example to show that the method of cross-validation needs modification when data are correlated

3.2 Proposed bandwidth selection criteria

As in Chapter 2, we assume that the data $(\boldsymbol{X}_i, Y_i)$ $(i = 1, \ldots, n)$ come from the model

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i.$$

In the random design case, the variance-covariance structure of the errors ε_i is associated with their locations. We assume that the random errors are second order stationary. That is,

$$\begin{aligned} \mathrm{E}(\varepsilon_i|\boldsymbol{X}_i) &= 0, \\ \mathrm{Var}(\varepsilon_i|\boldsymbol{X}_i) &= \sigma^2, \end{aligned}$$

$$\text{Cov}(\varepsilon_i, \varepsilon_j | \boldsymbol{X}_i, \boldsymbol{X}_j) \;=\; \sigma^2 \rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j).$$

Notice that the correlation function $\rho_n(\boldsymbol{x})$ varies with the sample size. As discussed in chapter 2, if the consistency of kernel smoothing estimators of the mean function is desired, $\rho_n(\boldsymbol{x})$ must shrink as the sample size n goes to ∞. In practical applications, however, there is only a fixed sample size. To simplify notation, we will later drop the subscript in $\rho_n(\boldsymbol{x})$.

Denote the correlation matrix as

$$\boldsymbol{\rho} = (\rho_n(\boldsymbol{X}_i - \boldsymbol{X}_j))_{n \times n} \equiv (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_n),$$

where the j-th column of $\boldsymbol{\rho}$ is

$$\boldsymbol{\rho}_j = (\rho_n(\boldsymbol{X}_1 - \boldsymbol{X}_j), \ldots, \rho_n(\boldsymbol{X}_n - \boldsymbol{X}_j))^T.$$

Denote the vector of responses as

$$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T,$$

the mean vector as

$$\boldsymbol{m} = (m_1, m_2, \ldots, m_n)^T \quad \text{with} \;\; m_i = m(\boldsymbol{X}_i),$$

and the estimator of the mean vector as

$$\hat{\boldsymbol{m}} = (\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_n)^T \quad \text{with} \;\; \hat{m}_i = \hat{m}(\boldsymbol{X}_i; \boldsymbol{H}).$$

The estimator $\hat{\boldsymbol{m}}$ is called a linear smoothing estimator if there is a matrix $\boldsymbol{S}$, which is not related to the response vector $\boldsymbol{Y}$, such that

$$\hat{\boldsymbol{m}} = \boldsymbol{S}\boldsymbol{Y}.$$

The matrix $\boldsymbol{S} = (s_{ij})_{n \times n}$ is called the smoothing matrix. Let

$$\boldsymbol{S} = \begin{pmatrix} s^T_{\boldsymbol{X}_1} \\ \vdots \\ s^T_{\boldsymbol{X}_n} \end{pmatrix}.$$

Here $\boldsymbol{s}_{\boldsymbol{X}_i}$ is referred to as the smoothing vector at the point $\boldsymbol{X}_i$. We have

$$\hat{m}_i = \boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{Y}.$$

The Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator are all linear smoothing estimators. For the Priestley-Chao estimator (under uniformly random design),

$$\boldsymbol{s}_{\boldsymbol{X}_i} = \left(\frac{A(\Omega)}{n} K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_i), \ldots, \frac{A(\Omega)}{n} K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{X}_i) \right)^T, \tag{3.1}$$

with $A(\Omega)$ as the Lebesgue measure of the design region Ω. For the Nadaraya-Watson estimator,

$$\boldsymbol{s}_{\boldsymbol{X}_i} = \left(\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_i)}{\sum_{j=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)}, \ldots, \frac{K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{X}_i)}{\sum_{j=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)} \right)^T. \tag{3.2}$$

For the local linear estimator,

$$\boldsymbol{s}_{\boldsymbol{X}_i} = \boldsymbol{e}_1^T (\boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \boldsymbol{X}_{\boldsymbol{X}_i})^{-1} \boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i}, \tag{3.3}$$

where

$$\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T \in I\!R^{d+1},$$

$$\boldsymbol{X}_{\boldsymbol{X}_i} = \begin{pmatrix} 1 & (\boldsymbol{X}_1 - \boldsymbol{X}_i)^T \\ \vdots & \vdots \\ 1 & (\boldsymbol{X}_n - \boldsymbol{X}_i)^T \end{pmatrix},$$

and

$$\boldsymbol{W}_{\boldsymbol{X}_i} = \mathrm{diag}(K_{\boldsymbol{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_i), \ldots, K_{\boldsymbol{H}}(\boldsymbol{X}_n - \boldsymbol{X}_i)).$$

For kernel type smoothing, selection of the bandwidth matrix is a very important problem. Our objective here is to select a bandwidth $\boldsymbol{H}$ such that the corresponding estimator of the mean function $\hat{m}(\boldsymbol{x}, \boldsymbol{H})$ is as close as possible to the mean function in the entire design region. Mathematically, we want to select the bandwidth matrix $\boldsymbol{H}$ to minimize the Mean Integrated Squared Error:

$$\begin{aligned}\text{MISE}(\boldsymbol{H}) &= \int \text{MSE}(\boldsymbol{x}, \boldsymbol{H}) w(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int \text{E}\left\{(\hat{m}(\boldsymbol{x}, \boldsymbol{H}) - m(\boldsymbol{x}))^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} w(\boldsymbol{x}) d\boldsymbol{x},\end{aligned}$$

where $w \geq 0$ is the weight function chosen by the user. One popular choice is to choose $w(\boldsymbol{x})$ as the design density $f(\boldsymbol{x})$. Under this scenario, a large sample approximation of MISE would be the mean average squared error across all of the design points, that is,

$$\begin{aligned}\text{MASE}(\boldsymbol{H}) &= \frac{1}{n}\sum_{i=1}^{n} \text{E}\left\{(\hat{m}(\boldsymbol{X}_i, \boldsymbol{H}) - m(\boldsymbol{X}_i))^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{E}\left\{(\hat{m}_i - m_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\}. \qquad (3.4)\end{aligned}$$

So we would want to choose a bandwidth $\boldsymbol{H}$ to minimize MASE($\boldsymbol{H}$).

The next problem is how to estimate MASE($\boldsymbol{H}$) from data. We will start by considering the sum of squared residuals $\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2$. Its conditional expectation

$$\begin{aligned}&\text{E}\left\{\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\ &= \text{E}\left\{\sum_{i=1}^{n}(Y_i - m_i)^2 + \sum_{i=1}^{n}(\hat{m}_i - m_i)^2 - 2\sum_{i=1}^{n}(Y_i - m_i)(\hat{m}_i - m_i) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\ &= n\sigma^2 + \sum_{i=1}^{n} \text{E}\left\{(\hat{m}_i - m_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} - 2\sum_{i=1}^{n} \text{Cov}(\hat{m}_i, Y_i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n).\end{aligned}$$

Note that

$$\text{Cov}(\hat{m}_i, Y_i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \boldsymbol{s}_{\boldsymbol{X}_i}^T \text{Cov}(\boldsymbol{Y}, Y_i | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \sigma^2 \boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i$$

Hence

$$\begin{aligned}\text{E}\left\{\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} &= n\sigma^2 + n\,\text{MASE}(\boldsymbol{H}) - 2\sigma^2 \sum_{i=1}^{n} \boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i \\ &= n\sigma^2 + n\,\text{MASE}(\boldsymbol{H}) - 2\sigma^2\,\text{tr}(\boldsymbol{S\rho}),\end{aligned}$$

which is equivalent to

$$\text{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} + \frac{2}{n}\sigma^2\,\text{tr}(\boldsymbol{S\rho}) = \sigma^2 + \text{MASE}(\boldsymbol{H}). \qquad (3.5)$$

Equation (3.5) implies that $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{m}_i)^2+\frac{2}{n}\sigma^2\ \mathrm{tr}(\boldsymbol{S\rho})$ is (conditionally) unbiased for $\sigma^2+\mathrm{MASE}(\boldsymbol{H})$, where σ^2 is a constant. It can be calculated when σ^2 and the correlation matrix $\boldsymbol{\rho}$ are known, So we have the following bandwidth selection criterion.

<u>**Criterion 1**</u>: *For a linear smoothing estimator, choose the bandwidth $\boldsymbol{H}$ by solving*

$$\min_{\boldsymbol{H}} \mathrm{C}_{\mathrm{L_c}}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{m}_i)^2+\frac{2}{n}\sigma^2\ \mathrm{tr}(\boldsymbol{S\rho}). \tag{3.6}$$

We call this criterion “the corrected C_L criterion”, C_{L_c}, because it generalizes Mallows C_L criterion to the case of correlated data by considering the correlations between errors. Note that this criterion is valid for the bandwidth selection of all linear smoothers, including the Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator.

For most linear smoothing estimators, the smoothing matrix $\boldsymbol{S}$ satisfies $\boldsymbol{S}\mathbf{1}_n = \mathbf{1}_n$, or equivalently, $\boldsymbol{s}_{\boldsymbol{X}_i}^T\mathbf{1}_n = 1$, $(i = 1,\dots,n)$. In other words, the sum of all the elements of the smoothing vector at any design point is equal to 1. Hence the estimator for $m(\boldsymbol{X}_i)$, $\boldsymbol{s}_{\boldsymbol{X}_i}^T\boldsymbol{Y}$, is a weighted average of the responses. For this type of linear smoothing estimator, the corrected C_L criterion has an equivalent form in terms of a semi-variogram. Following Cressie (1991), given two spatial locations $\boldsymbol{X}$ and $\boldsymbol{X}+\boldsymbol{d}$, the semi-variogram $\gamma(\boldsymbol{d})$ is

$$\gamma(\boldsymbol{d}) = \frac{1}{2}\mathrm{E}\left\{(\varepsilon(\boldsymbol{X}+\boldsymbol{d})-\varepsilon(\boldsymbol{X}))^2|\boldsymbol{X},\boldsymbol{d}\right\}. \tag{3.7}$$

The semi-variogram function $\gamma(\boldsymbol{d})$ has to be an even function, i.e., $\gamma(\boldsymbol{d}) = \gamma(-\boldsymbol{d})$. It can be easily seen that

$$\mathsf{Cov}(\varepsilon(\boldsymbol{X}_i),\varepsilon(\boldsymbol{X}_j)|\boldsymbol{X}_i,\boldsymbol{X}_j) = \sigma^2-\gamma(\boldsymbol{X}_i-\boldsymbol{X}_j).$$

This leads to a matrix relation between the correlation matrix $\boldsymbol{\rho}$ and the semi-variogram matrix $\boldsymbol{\Gamma} = (\gamma(\boldsymbol{X}_i-\boldsymbol{X}_j))_{n\times n}$:

$$\sigma^2\boldsymbol{\rho} = \sigma^2\mathbf{1}_n\ \mathbf{1}_n^T - \boldsymbol{\Gamma}. \tag{3.8}$$

When $\boldsymbol{S}\ \mathbf{1}_n = \mathbf{1}_n$, by multiplying $\boldsymbol{S}$ to the left of both sides of equation (3.8), we have

$$\sigma^2 \boldsymbol{S}\boldsymbol{\rho} = \sigma^2 \mathbf{1}_n\ \mathbf{1}_n^T - \boldsymbol{S}\boldsymbol{\Gamma}.$$

By taking the trace of both sides of the above equation, we get

$$\sigma^2 \mathrm{tr}(\boldsymbol{S}\boldsymbol{\rho}) = n\sigma^2 - \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Gamma}).$$

Thus

$$\mathrm{C_{L_c}}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 + 2\sigma^2 - \frac{2}{n}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Gamma}).$$

Hence we have the following criterion, which is equivalent to the "corrected C_L" criterion when $\boldsymbol{S}\ \mathbf{1}_n = \mathbf{1}_n$.

<u>**Criterion 1'**</u>: *For a linear smoothing estimator whose smoothing matrix satisfies $\boldsymbol{S}\ \mathbf{1}_n = \mathbf{1}_n$, choose the bandwidth $\boldsymbol{H}$ by solving*

$$\min_{\boldsymbol{H}} \mathrm{C'}_{L_c}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 - \frac{2}{n}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Gamma}). \tag{3.9}$$

Criterion 1' uses the semi-variogram matrix $\boldsymbol{\Gamma}$, while Criterion 1 uses the variance σ^2 and the correlation matrix $\boldsymbol{\rho}$. In spatial statistics, usually $\boldsymbol{\Gamma}$ can be estimated more efficiently than the covariance matrix $\sigma^2\boldsymbol{\rho}$ (Cressie (1991, pages 70-73)).

The Nadaraya-Watson estimator satisfies $\boldsymbol{S}\ \mathbf{1}_n = \mathbf{1}_n$, because equation (3.2) immediately leads to $\boldsymbol{s}_{\boldsymbol{X}_i}^T \mathbf{1}_n = 1$, $(i = 1, \ldots, n)$.

The local linear estimator also satisfies $\boldsymbol{S}\ \mathbf{1}_n = \mathbf{1}_n$. To see this, note that

$$\left(\boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \boldsymbol{X}_{\boldsymbol{X}_i}\right)^{-1} \left(\boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \boldsymbol{X}_{\boldsymbol{X}_i}\right) = \boldsymbol{I}_{d+1}.$$

Hence

$$\left(\boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \boldsymbol{X}_{\boldsymbol{X}_i}\right)^{-1} \boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \mathbf{1}_n = e_1,$$

which leads to

$$\boldsymbol{s}_{\boldsymbol{X}_i}^T \mathbf{1}_n = e_1^T \left(\boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \boldsymbol{X}_{\boldsymbol{X}_i}\right)^{-1} \boldsymbol{X}_{\boldsymbol{X}_i}^T \boldsymbol{W}_{\boldsymbol{X}_i} \mathbf{1}_n = 1$$

The Priestley-Chao estimator, however, only satisfies $\boldsymbol{S}\ \mathbf{1}_n \approx \mathbf{1}_n$.

So Criterion 1' can be applied to the Nadaraya-Watson estimator and the local linear estimator directly. Criterion 1' can also be applied to the Priestley-Chao estimator, but it is only approximately equivalent to Criterion 1.

When data are uncorrelated, Criterion 1 reduces to the Mallows C_L criterion:

$$\min_{\boldsymbol{H}} \mathrm{C}_L(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 + \frac{2}{n}\sigma^2\ \mathrm{tr}(\boldsymbol{S}). \tag{3.10}$$

This is the original Mallows C_L criterion multiplied by σ^2. To apply this criterion, one must estimate the variance of the errors. One way to avoid this estimation problem is cross-validation. The cross-validation criterion chooses the bandwidth matrix $\boldsymbol{H}$ by minimizing

$$\mathrm{CV}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_{i,(-i)}(\boldsymbol{X}_i;\boldsymbol{H})\right)^2, \tag{3.11}$$

where $\hat{m}_{i,(-i)}(\boldsymbol{X}_i;\boldsymbol{H})$ is the estimator of $m(\boldsymbol{X}_i)$ with bandwidth $\boldsymbol{H}$ without using the i-th datum $(\boldsymbol{X}_i,\ Y_i)$.

For linear smoothing estimators that satisfy $\boldsymbol{S}\ \mathbf{1}_n = \mathbf{1}_n$, it turns out that

$$\hat{m}_{i,(-i)}(\boldsymbol{X}_i;\boldsymbol{H}) = \frac{\hat{m}_i - s_{ii}Y_i}{1 - s_{ii}}, \tag{3.12}$$

where s_{ii} is the i-th entry of the smoothing vector $\boldsymbol{s}_{\boldsymbol{X}_i}$. By substituting (3.12) into (3.11), we get

$$\mathrm{CV}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_i}{1 - s_{ii}}\right)^2 \tag{3.13}$$

In the case of a fixed unequally spaced design or a random design, the generalized cross-validation (GCV), proposed by Craven and Wahba (1979), is often used. The GCV is defined as

$$\mathrm{GCV}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_i}{1 - \frac{1}{n}\sum_{i=1}^{n} s_{ii}}\right)^2. \tag{3.14}$$

When data are correlated, both CV and GCV are inappropriate for selecting the bandwidth. They need to be modified by considering the correlations between errors.

In the following discussion, we will use the assumptions $(A1) - (A5)$ in Chapter 2 and add the following assumption:

$(A6)$ $\lim_{n\to\infty} n \int \rho_n^2(\boldsymbol{t})d\boldsymbol{t} = \rho_{II}$.

Again, we suppose that Ω is a bounded open closure.

The following lemma is important for our discussion of the cross-validation type bandwidth selection.

Lemma 3.1 *Let $\boldsymbol{s}_{\boldsymbol{X}_i}$ be the smoothing vector of either the local linear estimator, the Priestley-Chao estimator, or the Nadaraya-Watson estimator (When the Priestley-Chao estimator is discussed, the design is assumed to be a uniformly random design. When the local linear estimator or the Nadaraya-Watson estimator is discussed, the design is assumed to be a general random design). Under Assumptions (A1)-(A6),*

$$\boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i = O_p(\frac{1}{n|\boldsymbol{H}|}), \tag{3.15}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i = \frac{K(\mathbf{0})(A(\Omega)+\rho_I)}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}). \tag{3.16}$$

The proof of Lemma 3.1 is presented in Section 3.4.

We now state the modified version of CV and GCV. The modified version of CV, called the corrected CV criterion, is given as follows.

<u>**Criterion 2**</u>: *For the local linear estimator, the Priestley-Chao estimator, or the Nadaraya-Watson estimator, choose the bandwidth $\boldsymbol{H}$ by solving*

$$\min_{\boldsymbol{H}} \mathrm{CV}_c(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_i}{1 - \boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i}\right)^2 \tag{3.17}$$

or the asymptotically equivalent form

$$\min_{\boldsymbol{H}} \mathrm{CV}_c'(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(1 + 2\boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i\right)(Y_i - \hat{m}_i)^2. \tag{3.18}$$

The modified version of GCV, called the corrected GCV criterion, is given as follows.

<u>Criterion 3</u>: *For the local linear estimator, the Priestley-Chao estimator, or the Nadaraya-Watson estimator, choose the bandwidth* $\boldsymbol{H}$ *by solving*

$$\min_{\boldsymbol{H}} \mathrm{GCV}_c(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{m}_i}{1-\frac{1}{n}\mathrm{tr}(\boldsymbol{S\rho})}\right)^2 \tag{3.19}$$

or the asymptotically equivalent form

$$\min_{\boldsymbol{H}} \mathrm{GCV}_c'(\boldsymbol{H}) = \frac{1}{n}\left(1+\frac{2}{n}\mathrm{tr}(\boldsymbol{S\rho})\right)\sum_{i=1}^{n}(Y_i - \hat{m}_i)^2 . \tag{3.20}$$

Note that $\mathrm{CV}_c(\boldsymbol{H})$ and $\mathrm{CV}_c'(\boldsymbol{H})$ (in Criterion 2), $\mathrm{GCV}_c(\boldsymbol{H})$ and $\mathrm{GCV}_c'(\boldsymbol{H})$ (in Criterion 3) require knowledge of the correlation matrix of the errors, but not the variance. They are asymptotically unbiased for $\sigma^2 + \mathrm{MASE}(\boldsymbol{H})$. Note also that $\mathrm{C}_{L_c}(\boldsymbol{H})$ (in Criterion 1) is unbiased for $\sigma^2 + \mathrm{MASE}(\boldsymbol{H})$, but require knowledge of the covariance matrix. So when the covariance matrix of the errors is known, Criterion 1 is preferred. But there are some situations in which Criterion 1 might be inferior, for example, the situation in which the variance and the correlations between the errors need to be estimated from data and the correlations can be more efficiently estimated than the variance.

Now we try to justify these criteria. By Taylor expansion, for a small number ε,

$$\frac{1}{(1-\varepsilon)^2} = 1 + 2\varepsilon + \frac{3}{(1-\theta\varepsilon)^4}\varepsilon^2 = 1 + 2\varepsilon + O(\varepsilon^2),$$

with $0 < \theta < 1$. Hence Lemma 3.1 implies that when the sample size is large,

$$\frac{1}{(1-\boldsymbol{s}_{\boldsymbol{X}_i}^T\rho_i)^2} = 1 + 2\boldsymbol{s}_{\boldsymbol{X}_i}^T\rho_i + o_p(\frac{1}{n|\boldsymbol{H}|}), \tag{3.21}$$

and

$$\frac{1}{(1-\frac{1}{n}\mathrm{tr}(\boldsymbol{S\rho}))^2} = 1 + \frac{2}{n}\mathrm{tr}(\boldsymbol{S\rho}) + o_p(\frac{1}{n|\boldsymbol{H}|}). \tag{3.22}$$

These two equations imply that $\mathrm{CV}_c'(\boldsymbol{H})$ and $\mathrm{GCV}_c'(\boldsymbol{H})$ are asymptotically equivalent to $\mathrm{CV}_c(\boldsymbol{H})$ and $\mathrm{GCV}_c(\boldsymbol{H})$ respectively.

In the following, we restrict our attention to the local linear estimator. The discussion of other estimators is analogous.

According to Theorem 2.1, under assumptions $(A1)$ to $(A5)$,

$$\begin{aligned}
&\mathrm{E}\left\{(\hat{m}_i - m_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\
&= \frac{1}{4}\mu_2^2(K)\mathrm{tr}^2(H^2\mathcal{H}_m(\boldsymbol{X}_i)) + \frac{\sigma^2\mu(K^2)(1 + f(\boldsymbol{X}_i)\ \rho_I)}{n|\boldsymbol{H}|f(\boldsymbol{X}_i)} + o_p\left(\lambda_{\max}^4(\boldsymbol{H}) + \frac{1}{n|\boldsymbol{H}|}\right).
\end{aligned}$$

Here $\lambda_{\max}^4(\boldsymbol{H})$ has the same convergence rate as $\mathrm{tr}^2(\boldsymbol{H}^2)$ when $n \to \infty$. Hence

$$\begin{aligned}
&\mathrm{MASE}(\boldsymbol{H}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{(\hat{m}_i - m_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\
&= \frac{1}{4}\mu_2^2(K) \int \mathrm{tr}^2(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{u}))f(\boldsymbol{u})d\boldsymbol{u} + \frac{\sigma^2\mu(K^2)(A(\Omega) + \rho_I)}{n|\boldsymbol{H}|} + o_p\left(\lambda_{\max}^4(\boldsymbol{H}) + \frac{1}{n|\boldsymbol{H}|}\right).
\end{aligned}$$

Notice that

$$\mathrm{E}\left\{\mathrm{GCV}_c(\boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} = \frac{\frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^n (Y_i - \hat{m}_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\}}{(1 - \frac{1}{n}\mathrm{tr}(\boldsymbol{S\rho}))^2}.$$

By (3.5) and (3.22),

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{GCV}_c(\boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\
&= \left(\sigma^2 + \mathrm{MASE}(\boldsymbol{H}) - \frac{2}{n}\sigma^2\mathrm{tr}(\boldsymbol{S\rho})\right)\left(1 + \frac{2}{n}\ \mathrm{tr}(\boldsymbol{S\rho}) + o_p(\frac{1}{n|\boldsymbol{H}|})\right) \\
&= \sigma^2 + \mathrm{MASE}(\boldsymbol{H}) + o_p\left(\lambda_M^4(\boldsymbol{H}) + \frac{1}{n|\boldsymbol{H}|}\right).
\end{aligned}$$

Hence the corrected GCV criterion is established.

Similar reasoning can be applied to the corrected CV criterion, with details omitted.

Next, we explain why the original GCV criterion fails when data are correlated.

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{GCV}(\boldsymbol{H}) | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} \\
&= \frac{\frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^n (Y_i - \hat{m}_i)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\}}{(1 - \frac{1}{n}\mathrm{tr}(\boldsymbol{S}))^2} \\
&= \left(\sigma^2 + \mathrm{MASE}(\boldsymbol{H}) - \frac{2}{n}\sigma^2\mathrm{tr}(\boldsymbol{S\rho})\right)\left(1 + \frac{2}{n}\mathrm{tr}(\boldsymbol{S}) + o_p(\frac{1}{n|\boldsymbol{H}|})\right)
\end{aligned}$$

$$= \sigma^2 + \frac{1}{4}\mu_2^2(K)\mathrm{tr}^2(\boldsymbol{H}^2\mathcal{H}_m(\boldsymbol{X}_i)) + \frac{\sigma^2\mu(K^2)}{n|\boldsymbol{H}|}\left\{A(\Omega) + \left(1 - \frac{2K(\mathbf{0})}{\mu(K^2)}\right)\rho_I\right\}$$
$$+o_p\left(\lambda_M^4(\boldsymbol{H}) + \frac{1}{n|\boldsymbol{H}|}\right).$$

For simplicity, we assume $\boldsymbol{H} = h\boldsymbol{H}_0$, with $\boldsymbol{H}_0$ a fixed matrix that does not change as the sample size n increases. Geometrically, $\boldsymbol{H}_0$ controls the shape of the elliptical neighborhood specified by the bandwidth matrix (such as the ratio of the axes of the ellipsis) and the orientation, while h controls its size. Then in this setup,

$$\mathrm{E}\left\{\mathrm{GCV}(\boldsymbol{H})|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\right\}$$
$$\approx \sigma^2 + \frac{1}{4}\mu_2^2(K)h^4\mathrm{tr}^2(\boldsymbol{H}_0^2\mathcal{H}_m(\boldsymbol{X}_i)) + \frac{\sigma^2\mu(K^2)}{nh^d|\boldsymbol{H}_0|}\left\{A(\Omega) + \left(1 - \frac{2K(\mathbf{0})}{\mu(K^2)}\right)\rho_I\right\}$$

When $2K(\mathbf{0}) > \mu(K^2)$, $1 - \frac{2K(\mathbf{0})}{\mu(K^2)}$ will be less than 0. If the correlation is positive and big enough such that

$$A(\Omega) + \left(1 - \frac{2K(\mathbf{0})}{\mu(K^2)}\right)\rho_I < 0, \qquad (3.23)$$

then the leading term of $\mathrm{E}\left\{\mathrm{GCV}(\boldsymbol{H})|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\right\}$ gets smaller when h decreases. This drives the bandwidth selected by GCV to 0 when positive correlations among observations are big enough.

The uniform spherical kernel and the Epanechnikov kernel both satisfy $2K(\mathbf{0}) > \mu(K^2)$. To see this, as an example, let us consider the Epanechnikov kernel:

$$K(\boldsymbol{x}) = \frac{d(d+2)}{2S_d}(1 - \|\boldsymbol{x}\|^2)\,1_{(\|\boldsymbol{x}\|<1)},$$

with S_d as the surface of the d-dimensional unit ball. For this kernel,

$$\frac{2K(\mathbf{0})}{\mu(K^2)} = \frac{d+4}{2} > 1.$$

So $2K(\mathbf{0}) > \mu(K^2)$. Hence for the Epanechnikov kernel, condition (3.23) becomes

$$A(\Omega) + \left(1 - \frac{d+4}{2}\right)\rho_I < 0,$$

In the simulated example of Section 3.1, $d = 2$, the Lebesgue measure of the design region $A(\Omega) = \mathbf{1}$, the sample size $n = 400$, and $\rho_n(\boldsymbol{x}) = \alpha^{20\|\boldsymbol{x}\|}$. So

$$\rho_I \approx n \int \rho_n(\boldsymbol{x}) d\boldsymbol{x} = 400(2\pi) \int_0^\infty \alpha^{20r} r dr = \frac{2\pi}{(\log \alpha)^2} \approx 4.335$$

Thus,

$$A(\Omega) + \left(1 - \frac{d+4}{2}\right) \rho_I \approx 1 + (1 - \frac{2+4}{2}) \times 4.335 = -7.670 < 0.$$

This illustrates why the GCV criterion tends to pick smaller smoothing parameters when data are positively correlated.

3.3 Simulation study

A simulation study is carried out to compare the classical bandwidth selection methods with the corrected versions proposed in Section 3.2. Only GCV, the corrected C_L, and the corrected GCV are considered.

Suppose that the design points $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ are uniformly sampled from the rectangle: $-1 < X_1 < 1$, $-1 < X_2 < 1$. The responses satisfy

$$Y_i = \sin(2\pi X_{i1}) + 4(X_{i2} - 0.5)^2 + \varepsilon_i,$$

where the random errors ε_i are normally distributed, with the covariance function

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j | \boldsymbol{X}_i, \boldsymbol{X}_j) = \sigma^2 \alpha^{20\|\boldsymbol{X}_i - \boldsymbol{X}_j\|}.$$

σ may take 2 possible values: 0.3 and 0.5. The signal to noise ratios (as defined in section 3.1) corresponding to these 2 values of σ are roughly 6.54:1 and 2.36:1, respectively. The parameter α, which controls the correlations between the errors, may take 3 values: 0 (no correlation), 0.3 (relatively small correlation), and 0.7 (relatively large correlation). The combination of the values of σ and the values of α leads to 6 different cases. For each case, 50 independent data sets of size 400 are simulated (different data sets have different sets of design points).

For each simulated data set, the local linear regression with the Epanechnikov kernel is then used to estimate the mean function. Two different approaches are used to search bandwidth matrices. The first approach is to consider diagonal bandwidths. For this approach, $\boldsymbol{H} = \mathrm{diag}(h_1, h_2)$, with h_1 and h_2 taking values from 0.1, 0.2, 0.3, 0.4, and 0.5. The second approach is to consider spherical bandwidths only. For this approach, the bandwidth matrix $\boldsymbol{H} = h\boldsymbol{I}_2$, with h taking values from 0.1, 0.2, 0.3, 0.4, and 0.5. In total, 25 bandwidth candidates are considered for the first "diagonal" approach, while 5 bandwidth candidates are considered for the second "spherical" approach. In this study, the covariance function is assumed completely known. For each simulated data set, each of the three criteria: GCV, the corrected C_L, and the corrected GCV, is used to pick a bandwidth from the available candidates. For purpose of comparison, we also calculate the "optimal" bandwidth. The "optimal" bandwidth can be obtained by picking the bandwidth from the available candidates that minimizes

$$\mathrm{E}\left\{(\hat{\boldsymbol{m}} - \boldsymbol{m})^T(\hat{\boldsymbol{m}} - \boldsymbol{m})|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\right\} = (\boldsymbol{Sm} - \boldsymbol{m})^T(\boldsymbol{Sm} - \boldsymbol{m}) + \sigma^2\mathrm{tr}(\boldsymbol{S\rho S}^T). \quad (3.24)$$

In the process of finding the "optimal" bandwidth, the true values of σ and α, and the true mean function are used. For each simulated data set, corresponding to every bandwidth selected by the 4 different methods, we calculate the average squared error using $\frac{1}{n}\sum_{i=1}^{n}(\hat{m}_i - m_i)^2$.

The bandwidths and average squared errors from 50 simulated data sets are summarized in Tables 3.1, 3.2, 3.3, and 3.4.

Table 3.1 displays the average values of (h_1, h_2), which are selected from the 25 diagonal bandwidth matrices using GCV, the corrected C_L, and the corrected GCV criteria, and the optimal selection. For every combination of σ and α, corresponding to every bandwidth selection criterion, the mean of the selected h_1's from the 50 simulated data sets and the mean of the selected h_2's from the 50 simulated data are listed in this table. When data are uncorrelated ($\alpha = 0$), essentially no difference can be found for the 3 criteria, regardless the magnitude of the noises. The reason is simple: when data are uncorrelated, the corrected GCV reduces to GCV, and the corrected C_L reduces to Mallows C_L, which is asymptotically equivalent to GCV. When data are correlated, GCV tends to select $h_1 = 0.1$ and $h_2 = 0.1$, corresponding to the bandwidth region of the smallest size. Compared to the bandwidths selected by GCV, the bandwidths selected by both the corrected C_L and the corrected GCV are much closer to the optimal bandwidths. If we fix the variance σ^2 and increase the correlation (increase α), or if we fix the correlation, increase the variance, the size of the bandwidth region selected by the corrected C_L bandwidths and the corrected GCV will increase.

Table 3.1 The average values of $(h_1,\ h_2)$ selected by 3 criteria, and the average values of optimal (h_1, h_2) (using diagonal bandwidths)

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	(0.1600, 0.2960)	(0.1580, 0.2900)	(0.1600, 0.2960)	(0.1780, 0.3000)
0.3	0.3	(0.1000, 0.1020)	(0.1900, 0.3580)	(0.1900, 0.3620)	(0.2000, 0.3480)
0.3	0.7	(0.1000, 0.1000)	(0.2040, 0.4120)	(0.1980, 0.4240)	(0.2000, 0.4000)
0.5	0	(0.1860, 0.3960)	(0.1840, 0.3960)	(0.1860, 0.3960)	(0.2000, 0.4000)
0.5	0.3	(0.1000, 0.1000)	(0.2740, 0.4360)	(0.2700, 0.4360)	(0.2040, 0.5000)
0.5	0.7	(0.1000, 0.1000)	(0.3540, 0.4640)	(0.3300, 0.4540)	(0.3000, 0.5000)

Table 3.2 displays the average values of h's based on 50 simulations by searching among the 5 spherical bandwidths (a much narrower search). The patterns shown in this table are similar to the patterns in Table 3.1. Again, no difference can be found among these criteria when data are uncorrelated. When correlations exist, GCV tends to select the smallest h. The corrected C_L and the corrected GCV select bandwidths much closer to the optimal bandwidths than GCV does.

Table 3.3 shows the means of average squared errors based on 50 simulations using the diagonal bandwidths selected by the 3 criteria. For each combination of σ and α, corresponding to every bandwidth selection criterion, the average squared error $\frac{1}{400}\sum_{i=1}^{400}(\hat{m}_i - m_i)^2$ for each simulated data set is calculated. Then the mean of the av-

Table 3.2 The average values of h selected by 3 criteria and the average values of optimal h (using spherical bandwidths only)

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.2000	0.2000	0.2000	0.2000
0.3	0.3	0.1000	0.2120	0.2140	0.2000
0.3	0.7	0.1000	0.2420	0.2460	0.2920
0.5	0	0.2500	0.2440	0.2500	0.2020
0.5	0.3	0.1000	0.3320	0.3380	0.3000
0.5	0.7	0.1000	0.4180	0.4280	0.4020

erage squared errors from the 50 simulations is calculated and listed in this table. When data are uncorrelated ($\alpha = 0$), the means of average squared errors are essentially equal for these 3 criteria. But when data are correlated ($\alpha = 0.3$ or 0.5), both the corrected C_L criterion and the corrected GCV criterion outperform the GCV criterion. The means of average squared errors for the corrected C_L and the corrected GCV criteria are very close to the results for the optimal bandwidths.

Table 3.3 The mean average squared errors using the diagonal bandwidths selected by 3 criteria and using the optimal diagonal bandwidths

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.0093	0.0093	0.0093	0.0090
0.3	0.3	0.0372	0.0236	0.0240	0.0225
0.3	0.7	0.0675	0.0552	0.0552	0.0515
0.5	0	0.0213	0.0209	0.0213	0.0196
0.5	0.3	0.1006	0.0455	0.0454	0.0416
0.5	0.7	0.1740	0.1168	0.1177	0.1104

Table 3.4 shows the means of average squared errors by searching spherical bandwidths only. Similar patterns as in Table 3.3 can be found. As in the search of the diagonal bandwidths, the 3 criteria are similar when data are uncorrelated. When data are correlated, the corrected C_L and the corrected GCV beat GCV for every combination of σ and α. The corrected C_L and the corrected GCV have similarly good performance.

If we compare corresponding cases between Table 3.3 and Table 3.4, for $\sigma = 0.3$ (relatively smaller noise), the diagonal bandwidth search does not provide a smaller

Table 3.4 The mean average squared error using spherical bandwidths selected by 3 criteria, and using the optimal spherical bandwidths

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.0127	0.0129	0.0127	0.0124
0.3	0.3	0.0373	0.0230	0.0234	0.0230
0.3	0.7	0.0675	0.0541	0.0541	0.0541
0.5	0	0.0207	0.0207	0.0207	0.0207
0.5	0.3	0.1006	0.0516	0.0520	0.0494
0.5	0.7	0.1740	0.1265	0.1275	0.1240

mean average squared error than the spherical bandwidth search in this study. For $\sigma = 0.5$ (relatively larger noise), when data are correlated, the broader diagonal bandwidth search does give a better result, but at the cost of more computing time.

The simulation study provides support for our proposed bandwidth selection criteria. The corrected GCV criterion and the corrected C_L criterion provide very similar results. The simulation study shows the benefit in kernel type regression by considering the correlation of the errors when data are actually correlated. In this study, we do not find significant benefit of conducting a broader diagonal bandwidth search. Thus a spherical bandwidth search may be good enough for this particular study.

Finally, we investigate the possibility of applying some of the asymptotic formulas derived in Chapter 2 for bandwidth selection. We use the above simulation study as an example. For a diagonal bandwidth $\boldsymbol{H} = \text{diag}(h_1, h_2)$, the AMISE (asymptotic mean integrated square error) formula (2.56) reduces to

$$\begin{aligned}
&\text{AMISE}(\boldsymbol{H}) \\
&= \frac{1}{4}\mu_2^2(K)h_1^4 \int (\frac{\partial^2 m}{\partial x_1^2})^2 w(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{4}\mu_2^2(K)h_2^4 \int (\frac{\partial^2 m}{\partial x_2^2})^2 w(\boldsymbol{x})d\boldsymbol{x} \\
&\quad + \frac{1}{2}\mu_2^2(K)h_1^2h_2^2 \int \frac{\partial^2 m}{\partial x_1^2}\frac{\partial^2 m}{\partial x_2^2} w(\boldsymbol{x})d\boldsymbol{x} + \frac{\sigma^2\mu(K^2)\,(\int w(\boldsymbol{x})/f(\boldsymbol{x})d\boldsymbol{x} + \rho_I)}{nh_1h_2}.
\end{aligned}$$

In the simulation study, the sample size $n = 400$, and the density function $f(\boldsymbol{x}) = 1$. We choose the weight function as $w(\boldsymbol{x}) = f(\boldsymbol{x}) = 1$. For the bivariate Epanechnicov kernel K, $\mu_2(K) = 1/6$, and $\mu(K^2) = \frac{4}{3\pi}$. For the mean function $m(\boldsymbol{x}) = \sin(2\pi x_1) +$

$4(x_2 - 0.5)^2$, $\int(\frac{\partial^2 m}{\partial x_1^2})^2 w(\boldsymbol{x})d\boldsymbol{x} = 8\pi^4$, $\int(\frac{\partial^2 m}{\partial x_2^2})^2 w(\boldsymbol{x})d\boldsymbol{x} = 64$, and $\int \frac{\partial^2 m}{\partial x_1^2}\frac{\partial^2 m}{\partial x_2^2} w(\boldsymbol{x})d\boldsymbol{x} = 0$. For the correlation $\rho_n(\boldsymbol{x}) = \alpha^{20\|\boldsymbol{x}\|}$, $\rho_I = n\int \alpha^{20\|\boldsymbol{x}\|}d\boldsymbol{x} = \frac{2\pi}{\log^2 \alpha}$. So

$$\text{AMISE}(\boldsymbol{H}) = \frac{\pi^4}{18}h_1^4 + \frac{4}{9}h_2^4 + \frac{\sigma^2(1 + 2\pi/(\log^2 \alpha))}{300\pi h_1 h_2}.$$

According to the above formula, for each of the 6 cases based on the 2 values of σ and the 3 values of α, the diagonal bandwidth from the 25 available candidates that minimizes AMISE can be calculated. Table 3.5 displays the average values of optimal (h_1, h_2) of the 50 simulations and the calculated AMISE-optimal bandwidths.

Table 3.5 The average values of optimal (h_1, h_2) from the 50 simulations and the AMISE-optimal bandwidths by searching the 25 diagonal bandwidths.

σ	α	Optimal	AMISE-optimal
0.3	0	(0.1780, 0.3000)	(0.1, 0.2)
0.3	0.3	(0.2000, 0.3480)	(0.2, 0.3)
0.3	0.7	(0.2000, 0.4000)	(0.2, 0.4)
0.5	0	(0.2000, 0.4000)	(0.1, 0.3)
0.5	0.3	(0.2040, 0.5000)	(0.2, 0.3)
0.5	0.7	(0.3000, 0.5000)	(0.3, 0.5)

We see that for some cases, the AMISE-optimal bandwidths are not reliable and away from the optimal bandwidths. Notice that when calculating the AMISE-optimal bandwidths, we have assumed that the expression of the true mean function m is known. In practice, the information about m has to be estimated from data through some procedures (say "plug-in" technique), so that the bandwidth selection using the AMISE formula becomes more difficult. Based on Table 3.5 and Table 3.1, we recommend other criteria, such as the corrected C_L criterion and the corrected GCV criterion, for bandwidth selection, particularly when the sample size is not very big, say less than 1000.

3.4 Technical proof

In this section, we will prove Lemma 3.1. First, we introduce a proposition.

Proposition 3.1 *Suppose that $q_i(\boldsymbol{u})$ $(i = 1, 2, 3)$ are continuous in the design region Ω, a bounded open closure, and $|q_i(\cdot)|$ $(i = 1, 2, 3)$ are bounded above. Suppose that $g(\boldsymbol{u}, \boldsymbol{v})$ is continuous in $\Omega \times \Omega$, and $|g(\boldsymbol{u}, \boldsymbol{v})|$ is bounded above. Then*

$$\lim_{n\to\infty} n|\boldsymbol{H}| \int\int K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho_n(\boldsymbol{u}-\boldsymbol{v})g(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} = K(\boldsymbol{0})\rho_I \int g(\boldsymbol{v},\boldsymbol{v})d\boldsymbol{v}, \qquad (3.25)$$

$$\lim_{n\to\infty} n|\boldsymbol{H}|^2 \int\int K^2_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho^2_n(\boldsymbol{u}-\boldsymbol{v})g(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} = K^2(\boldsymbol{0})\ \rho_{II} \int g(\boldsymbol{v},\boldsymbol{v})d\boldsymbol{v}. \qquad (3.26)$$

and

$$\begin{aligned}\lim_{n\to\infty} n^2|\boldsymbol{H}|^2 \int\int\int & K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{w})\rho_n(\boldsymbol{u}-\boldsymbol{w})K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{w})\rho_n(\boldsymbol{v}-\boldsymbol{w})q_1(\boldsymbol{u})q_2(\boldsymbol{v}) \\ & \cdot q_3(\boldsymbol{w})d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w} \\ = \ & K^2(0)\rho_I^2 \int q_1(\boldsymbol{w})q_2(\boldsymbol{w})q_3(\boldsymbol{w})d\boldsymbol{w}. \qquad (3.27)\end{aligned}$$

Here all the integrals are over Ω.

Proof: The proof of this proposition is similar to the proof of Proposition 2.3 of Chapter 2. Let

$$g_n(\boldsymbol{v}) = n|\boldsymbol{H}| \int K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho_n(\boldsymbol{u}-\boldsymbol{v})g(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}.$$

Then

$$n|\boldsymbol{H}| \int\int K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho_n(\boldsymbol{u}-\boldsymbol{v})g(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} = \int g_n(\boldsymbol{v})d\boldsymbol{v}. \qquad (3.28)$$

By taking transformation: $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{H}\boldsymbol{t}$, we have

$$g_n(\boldsymbol{v}) = n|\boldsymbol{H}| \int K(\boldsymbol{t})\rho_n(\boldsymbol{H}\boldsymbol{t})g(\boldsymbol{v}+\boldsymbol{H}\boldsymbol{t},\boldsymbol{v})d\boldsymbol{t}.$$

Let $K_M = \max_\Omega(K(\boldsymbol{t}))$, and $g_M = \max_\Omega(g(\boldsymbol{u},\boldsymbol{v}))$. Then

$$\begin{aligned}|g_n(\boldsymbol{v})| &\le K_M g_M \left\{ n|\boldsymbol{H}| \int \rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t} \right\} \\ &\le K_M g_M \left\{ n \int |\rho_n(\boldsymbol{t})|d\boldsymbol{t} \right\}.\end{aligned}$$

By assumption (A4), $n\int|\rho_n(\boldsymbol{t})|d\boldsymbol{t}$ is bounded. Hence $g_n(\boldsymbol{v})$ is bounded. Because of the continuity of $K(\cdot)$ and $g(\cdot,\cdot)$, $g_n(\boldsymbol{v})$ is continuous.

For any fixed interior point $\boldsymbol{v}$, we can find a small enough $\varepsilon > 0$, such that the ball centered at $\boldsymbol{v}$ with radius ϵ, $\{\boldsymbol{u} : \|\boldsymbol{u}-\boldsymbol{v}\| < \varepsilon\}$, is contained in Ω. So,

$$\begin{aligned} g_n(\boldsymbol{v}) &= n|\boldsymbol{H}|\int\{K(\mathbf{0})+(K(\boldsymbol{t})-K(\mathbf{0}))\}\rho_n(\boldsymbol{H}\boldsymbol{t})\{g(\boldsymbol{v},\boldsymbol{v})+o(1)\}\,d\boldsymbol{t} \\ &= K(\mathbf{0})g(\boldsymbol{v},\boldsymbol{v})n|\boldsymbol{H}|\int\rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t}+g(\boldsymbol{v},\boldsymbol{v})n|\boldsymbol{H}|\int_{\|\boldsymbol{t}\|<\varepsilon}(K(\boldsymbol{t})-K(\mathbf{0}))\rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t} \\ &\quad +g(\boldsymbol{v},\boldsymbol{v})n|\boldsymbol{H}|\int_{\|\boldsymbol{t}\|\geq\varepsilon}(K(\boldsymbol{t})-K(\mathbf{0}))\rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t}+o(1) \\ &\equiv I_1+I_2+I_3+o(1). \end{aligned} \tag{3.29}$$

By assumption (A3),

$$I_1 = K(\mathbf{0})g(\boldsymbol{v},\boldsymbol{v})n|\boldsymbol{H}|\int\rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t} = K(\mathbf{0})g(\boldsymbol{v},\boldsymbol{v})n\int\rho_n(\boldsymbol{t})d\boldsymbol{t} = K(\mathbf{0})\;\rho_I\;g(\boldsymbol{v},\boldsymbol{v}). \tag{3.30}$$

Since the kernel function is Lipschitz continuous (Assumption (A2) in Chapter 2),

$$\begin{aligned} |I_2| &= |g(\boldsymbol{v},\boldsymbol{v})n|\boldsymbol{H}|\int_{\|\boldsymbol{t}\|<\varepsilon}(K(\boldsymbol{t})-K(\mathbf{0}))\rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t}| \\ &\leq g_M L\varepsilon n|\boldsymbol{H}|\int_{\|\boldsymbol{t}\|<\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{t})|d\boldsymbol{t}. \end{aligned}$$

By assumption (A4),

$$\int_{\|\boldsymbol{t}\|<\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{t})|d\boldsymbol{t} \leq C.$$

So

$$|I_2| \leq g_M LC\varepsilon. \tag{3.31}$$

Then consider the I_3,

$$\begin{aligned} |I_3| &= |g(\boldsymbol{v},\boldsymbol{v})n|\boldsymbol{H}|\int_{\|\boldsymbol{t}\|\geq\varepsilon}(K(\boldsymbol{t})-K(\mathbf{0}))\rho_n(\boldsymbol{H}\boldsymbol{t})d\boldsymbol{t}| \\ &\leq g_M K_M n|\boldsymbol{H}\int_{\|\boldsymbol{t}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{t})|d\boldsymbol{t}. \end{aligned}$$

By proposition 2.1, $\lim_{n\to\infty}n|\boldsymbol{H}|\int_{\|\boldsymbol{t}\|\geq\varepsilon}|\rho_n(\boldsymbol{H}\boldsymbol{t})|d\boldsymbol{t}=0$. So,

$$\lim_{n\to\infty}|I_3| = 0. \tag{3.32}$$

Since ε can be arbitrarily small, we have

$$\lim_{n\to\infty} g_n(\boldsymbol{v}) = K(\mathbf{0})\ \rho_I\ g(\boldsymbol{v},\boldsymbol{v}). \tag{3.33}$$

Now, we have shown that $g_n(\boldsymbol{v})$ is bounded, continuous, and converges to $K(\mathbf{0})\ \rho_I\ g(\boldsymbol{v},\boldsymbol{v})$ in the interior of Ω. Also notice that the set Ω has a finite measure. Hence by Lebesgue bounded convergence theorem,

$$\lim_{n\to\infty} \int g_n(\boldsymbol{v})d\boldsymbol{v} = \int \lim_{n\to\infty} g_n(\boldsymbol{v})d\boldsymbol{v} = K(\mathbf{0})\ \rho_I\ \int g(\boldsymbol{v},\boldsymbol{v})d\boldsymbol{v}.$$

This completes the proof of equation (3.25). The proofs of (3.26) and (3.27) are similar. We only sketch the ideas as follows.

To prove (3.26), let

$$g_{1n}(\boldsymbol{v}) = n|\boldsymbol{H}|^2 \int K^2_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho_n^2(\boldsymbol{u}-\boldsymbol{v})g(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}.$$

Then

$$n|\boldsymbol{H}|^2 \int\int K^2_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho_n^2(\boldsymbol{u}-\boldsymbol{v})g(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} = \int g_{1n}(\boldsymbol{v})d\boldsymbol{v}. \tag{3.34}$$

We can show that for any interior point $\boldsymbol{v}$ in Ω,

$$\lim_{n\to\infty} g_{1n}(\boldsymbol{v}) = K^2(\mathbf{0})\ \rho_{II}\ g(\boldsymbol{v},\boldsymbol{v}). \tag{3.35}$$

Then using the Lebesgue bounded convergence theorem, we have

$$\lim_{n\to\infty} \int g_{1n}(\boldsymbol{v})d\boldsymbol{v} = \int \lim_{n\to\infty} g_{1n}(\boldsymbol{v})d\boldsymbol{v} = K^2(\mathbf{0})\ \rho_{II}\ \int g(\boldsymbol{v},\boldsymbol{v})d\boldsymbol{v}. \tag{3.36}$$

This completes the proof of (3.26).

To prove (3.27), let

$$q_{1n}(\boldsymbol{w}) = n|\boldsymbol{H}| \int K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{w})\rho_n(\boldsymbol{u}-\boldsymbol{w})q_1(\boldsymbol{u})d\boldsymbol{u},$$

and

$$q_{2n}(\boldsymbol{w}) = n|\boldsymbol{H}| \int K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{w})\rho_n(\boldsymbol{v}-\boldsymbol{w})q_2(\boldsymbol{u})d\boldsymbol{v}.$$

Then

$$\begin{aligned} & n^2|\boldsymbol{H}|^2 \int\int\int K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{w})\rho_n(\boldsymbol{u}-\boldsymbol{w})K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{w})\rho_n(\boldsymbol{v}-\boldsymbol{w})q_1(\boldsymbol{u})q_2(\boldsymbol{v}) \\ & \quad \cdot q_3(\boldsymbol{w})d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w} \\ = & \int q_{1n}(\boldsymbol{w})q_{2n}(\boldsymbol{w})q_3(\boldsymbol{w})d\boldsymbol{w}. \end{aligned}$$

We can show that for any interior point $\boldsymbol{w}$ in Ω,

$$\lim_{n\to\infty} h_{in}(\boldsymbol{w}) = K(\boldsymbol{0})\ \rho_I\ q_i(\boldsymbol{w}) \ \ (i=1,2). \tag{3.37}$$

Again, use Lebesgue bounded convergence theorem, we have

$$\begin{aligned} \lim_{n\to\infty}\int q_{1n}(\boldsymbol{w})q_{2n}(\boldsymbol{w})h_3(\boldsymbol{w})d\boldsymbol{w} &= \int \lim_{n\to\infty} q_{1n}(\boldsymbol{w}) \lim_{n\to\infty} q_{2n}(\boldsymbol{w})q_3(\boldsymbol{w})d\boldsymbol{w} \\ &= K^2(\boldsymbol{0})\ \rho_I^2 \int q_1(\boldsymbol{w})q_2(\boldsymbol{w})q_3(\boldsymbol{w})d\boldsymbol{w}. \end{aligned}$$

This completes the proof of (3.27).

■

Now we are ready to prove Lemma 3.1. We consider the cases when the smoothing vector $\boldsymbol{S}_{\boldsymbol{X}_i}$ is from the local linear estimator, the Priestley-Chao estimator, and the Nadaraya-Watson estimator.

Proof: For the local linear estimator (the most difficult case),

$$\boldsymbol{s}_{\boldsymbol{X}_i}^T\boldsymbol{\rho}_i = \boldsymbol{e}_1^T(\frac{1}{n}\boldsymbol{X}_{\boldsymbol{X}_i}^T\boldsymbol{W}_{\boldsymbol{X}_i}\boldsymbol{X}_{\boldsymbol{X}_i})^{-1}(\frac{1}{n}\boldsymbol{X}_{\boldsymbol{X}_i}^T\boldsymbol{W}_{\boldsymbol{X}_i}\boldsymbol{\rho}_i).$$

From Ruppert and Wand (1994),

$$\boldsymbol{e}_1^T(\frac{1}{n}\boldsymbol{X}_{\boldsymbol{X}_i}^T\boldsymbol{W}_{\boldsymbol{X}_i}\boldsymbol{X}_{\boldsymbol{X}_i})^{-1} = (f^{-1}(\boldsymbol{X}_i)+o_p(1),\ \ -f^{-2}(\boldsymbol{X}_i)(\nabla_f(\boldsymbol{X}_i))^T + o_p(\boldsymbol{1}^T)).$$

Note that

$$\frac{1}{n}\boldsymbol{X}_{\boldsymbol{X}_i}^T\boldsymbol{W}_{\boldsymbol{X}_i}\boldsymbol{\rho}_i = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j-\boldsymbol{X}_i) \\ \frac{1}{n}\sum_{j=1}^n K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j-\boldsymbol{X}_i)(\boldsymbol{X}_j-\boldsymbol{X}_i) \end{pmatrix}.$$

We have

$$s_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i \approx \frac{1}{n}\sum_{j=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)/f(\boldsymbol{X}_i) \equiv T_n(\boldsymbol{X}_i). \tag{3.38}$$

We can show that for the Nadaraya-Watson estimator, the above equation holds as well, and for the Priestley-Chao estimator, the above equation becomes an equality.

Next, we consider the moments of $T_n(\boldsymbol{X}_i)$. Note that

$$T_n(\boldsymbol{X}_i) = \frac{K(\mathbf{0})}{n|\boldsymbol{H}|f(\boldsymbol{X}_i)} + \frac{1}{n}\sum_{j\neq i} K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)/f(\boldsymbol{X}_i). \tag{3.39}$$

Then,

$$\mathrm{E}(T_n(\boldsymbol{X}_i)) = \frac{K(\mathbf{0})A(\Omega)}{n|\boldsymbol{H}|} + (1 - 1/n)\int\int K_{\boldsymbol{H}}(\boldsymbol{v} - \boldsymbol{u})\rho_n(\boldsymbol{v} - \boldsymbol{u})f(\boldsymbol{u})d\boldsymbol{u}d\boldsymbol{v}. \tag{3.40}$$

By applying the first equation of Proposition 3.1,

$$n|\boldsymbol{H}|\int\int K_{\boldsymbol{H}}(\boldsymbol{u} - \boldsymbol{v})\rho_n(\boldsymbol{u} - \boldsymbol{v})f(\boldsymbol{u})d\boldsymbol{u}d\boldsymbol{v} = K(0)\ \rho_I + o(1).$$

Hence

$$\mathrm{E}(T_n(\boldsymbol{X}_i)) = \frac{K(\mathbf{0})(A(\Omega) + \rho_I)}{n|\boldsymbol{H}|} + o(\frac{1}{n|\boldsymbol{H}|}). \tag{3.41}$$

Note that

$$\begin{aligned}
& T_n^2(\boldsymbol{X}_i) \\
= \ & \frac{K^2(\mathbf{0})}{n^2|\boldsymbol{H}|^2 f^2(\boldsymbol{X}_i)} + 2\frac{K(\mathbf{0})}{n^2|\boldsymbol{H}|}\sum_{j\neq i} K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)/f^2(\boldsymbol{X}_i) \\
& + \frac{1}{n^2}\sum_{j\neq i}\sum_{l\neq i} K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)K_{\boldsymbol{H}}(\boldsymbol{X}_l - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_l - \boldsymbol{X}_i)/f^2(\boldsymbol{X}_i) \\
= \ & \frac{K^2(\mathbf{0})}{n^2|\boldsymbol{H}|^2 f^2(\boldsymbol{X}_i)} + 2\frac{K(\mathbf{0})}{n^2|\boldsymbol{H}|}\sum_{j\neq i} K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)/f^2(\boldsymbol{X}_i) \\
& + \frac{1}{n^2}\sum_{j\neq i} K_{\boldsymbol{H}}^2(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n^2(\boldsymbol{X}_j - \boldsymbol{X}_i)/f^2(\boldsymbol{X}_i) \\
& + \frac{1}{n^2}\sum_{j\neq i}\sum_{l\neq i,j} K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)K_{\boldsymbol{H}}(\boldsymbol{X}_l - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_l - \boldsymbol{X}_i)/f^2(\boldsymbol{X}_i).
\end{aligned}$$

By taking the expectation, we get

$$\begin{aligned}
&\mathrm{E}(T_n^2(\boldsymbol{X}_i)) \\
= \ & \frac{K^2(\mathbf{0})}{n^2|\boldsymbol{H}|^2}\int 1/f(\boldsymbol{u})d\boldsymbol{u} + 2\frac{K(\mathbf{0})(n-1)}{n^2|\boldsymbol{H}|}\int\int K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{u})\rho_n(\boldsymbol{v}-\boldsymbol{u})\frac{f(\boldsymbol{v})}{f(\boldsymbol{u})}d\boldsymbol{u}d\boldsymbol{v} \\
&+\frac{n-1}{n^2}\int\int K_{\boldsymbol{H}}^2(\boldsymbol{v}-\boldsymbol{u})\rho_n^2(\boldsymbol{v}-\boldsymbol{u})\frac{f(\boldsymbol{v})}{f(\boldsymbol{u})}d\boldsymbol{u}d\boldsymbol{v} \\
&+\frac{(n-1)(n-2)}{n^2}\int\int K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{u})\rho_n(\boldsymbol{v}-\boldsymbol{u})K_{\boldsymbol{H}}(\boldsymbol{w}-\boldsymbol{u})\rho_n(\boldsymbol{w}-\boldsymbol{u}) \\
&\cdot\frac{f(\boldsymbol{v})f(\boldsymbol{w})}{f(\boldsymbol{u})}d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w}.
\end{aligned}$$

Now applying Proposition 3.1 to the above double integrals, we have

$$\begin{aligned}
&\mathrm{E}(T_n^2(\boldsymbol{X}_i)) \\
= \ & \frac{K^2(\mathbf{0})}{n^2|\boldsymbol{H}|^2}\int 1/f(\boldsymbol{u})d\boldsymbol{u} + 2\frac{K(\mathbf{0})(n-1)}{n^3|\boldsymbol{H}|^2}(K(\mathbf{0})\ \rho_I A(\Omega)+o(1)) \\
&+\frac{n-1}{n^3|\boldsymbol{H}|^2}(K^2(\mathbf{0})\ \rho_{II}\ A(\Omega)+o(1)) + \frac{(n-1)(n-2)}{n^4|\boldsymbol{H}|^2}(K^2(\mathbf{0})\ \rho_{\cdot I}^{2}+o(1)) \\
\equiv \ & \frac{c_1}{n^2|\boldsymbol{H}|^2} + o(\frac{1}{n^2|\boldsymbol{H}|^2}). \qquad (3.42)
\end{aligned}$$

Then,

$$\mathrm{Var}(T_n(\boldsymbol{X}_i)) = \mathrm{E}(T_n^2(\boldsymbol{X}_i)) - \mathrm{E}^2(T_n(\boldsymbol{X}_i)) = O(\frac{1}{n^2|\boldsymbol{H}|^2}). \qquad (3.43)$$

Equations (3.41) and (3.43) imply

$$T_n(\boldsymbol{X}_i) = O(\frac{1}{n|\boldsymbol{H}|}).$$

Since the leading term of $\boldsymbol{s}_{\boldsymbol{X}_i}^T\rho_i$ is $T_n(\boldsymbol{X}_i)$,

$$\boldsymbol{s}_{\boldsymbol{X}_i}^T\rho_i = O(\frac{1}{n|\boldsymbol{H}|}).$$

This completes the proof of equation (3.15) of Lemma 3.1.

Next, consider the covariance between $T_n(\boldsymbol{X}_i)$ and $T_n(\boldsymbol{X}_j)$ (for $i \neq j$). Notice that

$$\begin{aligned}
T_n(\boldsymbol{X}_i) \ = \ & \frac{K(\mathbf{0})}{n|\boldsymbol{H}|f(\boldsymbol{X}_i)} + \frac{1}{n}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)} \\
&+\frac{1}{n}\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)} \\
\equiv \ & T_{n1}(i)+T_{n2}(i)+T_{n3}(i),
\end{aligned}$$

and

$$\begin{aligned} T_n(\boldsymbol{X}_j) &= \frac{K(\mathbf{0})}{n|\boldsymbol{H}|f(\boldsymbol{X}_j)} + \frac{1}{n}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)}{f(\boldsymbol{X}_j)} \\ &\quad + \frac{1}{n}\sum_{t \neq i,j} \frac{K_{\boldsymbol{H}}(\boldsymbol{X}_t - \boldsymbol{X}_j)\rho_n(\boldsymbol{X}_t - \boldsymbol{X}_i)}{f(\boldsymbol{X}_j)} \\ &\equiv T_{n1}(j) + T_{n2}(j) + T_{n3}(j). \end{aligned}$$

Then,

$$\begin{aligned} & \mathrm{Cov}(T_n(\boldsymbol{X}_i), T_n(\boldsymbol{X}_j)) \\ =\ & \mathrm{E}(T_n(\boldsymbol{X}_i)T_n(\boldsymbol{X}_j)) - \mathrm{E}(T_n(\boldsymbol{X}_i))\mathrm{E}(T_n(\boldsymbol{X}_j)) \\ =\ & \mathrm{E}(T_{n1}(i)T_{n1}(j)) + \mathrm{E}(T_{n1}(i)T_{n2}(j)) + \mathrm{E}(T_{n1}(i)T_{n3}(j)) \\ & + \mathrm{E}(T_{n2}(i)T_{n1}(j)) + \mathrm{E}(T_{n2}(i)T_{n2}(j)) + \mathrm{E}(T_{n1}(i)T_{n3}(j)) \\ & + \mathrm{E}(T_{n3}(i)T_{n1}(j)) + \mathrm{E}(T_{n3}(i)T_{n2}(j)) + \mathrm{E}(T_{n3}(i)T_{n3}(j)) \\ & - \mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n1}(j)) - \mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n2}(j)) - \mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n3}(j)) \\ & - \mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n1}(j)) - \mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n2}(j)) - \mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n2}(j)) \\ & - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n1}(j)) - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n2}(j)) - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n2}(j)). \end{aligned} \tag{3.44}$$

From the independence of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, it can be seen that

$$\mathrm{E}(T_{n1}(i)T_{n1}(j)) - \mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n1}(j)) = 0, \tag{3.45}$$

$$\mathrm{E}(T_{n1}(i)T_{n3}(j)) - \mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n3}(j)) = 0, \tag{3.46}$$

and

$$\mathrm{E}(T_{n3}(i)T_{n1}(j)) - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n1}(j)) = 0. \tag{3.47}$$

Since

$$\begin{aligned} \mathrm{E}(T_{n1}(i)T_{n2}(j)) &= E\left(\frac{K(\mathbf{0})}{n^2|\boldsymbol{H}|}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_j - \boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j - \boldsymbol{X}_i)}{f(\boldsymbol{X}_i)f(\boldsymbol{X}_j)}\right) \\ &= \frac{K(\mathbf{0})}{n^3|\boldsymbol{H}|^2}\left\{n|\boldsymbol{H}|\int\int K_{\boldsymbol{H}}(\boldsymbol{u}-\boldsymbol{v})\rho_n(\boldsymbol{u}-\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}\right\} \\ &= \frac{K^2(\mathbf{0})\ \rho_I A(\Omega)}{n^3|\boldsymbol{H}|^2} + o(\frac{1}{n^3|\boldsymbol{H}|^2})\ \text{(by Proposition 3.1)}, \end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n2}(j)) &= E\left(\frac{K(\mathbf{0})}{n^2|\boldsymbol{H}|f(\boldsymbol{X}_i)}\right) E\left(\frac{1}{n}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j-\boldsymbol{X}_i)}{f(\boldsymbol{X}_j)}\right) \\
&= \frac{K(\mathbf{0})\ \rho_I A(\Omega)}{n^3|\boldsymbol{H}|^2}\left\{n|\boldsymbol{H}|\int\int K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{u})\rho_n(\boldsymbol{v}-\boldsymbol{u})f(\boldsymbol{u})d\boldsymbol{u}d\boldsymbol{v}\right\} \\
&= \frac{K^2(\mathbf{0})\ \rho_I A(\Omega)}{n^3|\boldsymbol{H}|^2} + o(\frac{1}{n^3|\boldsymbol{H}|^2})\ \text{(by Proposition 3.1)},
\end{aligned}$$

we get

$$\mathrm{E}(T_{n1}(i)T_{n2}(j)) - \mathrm{E}(T_{n1}(i))\mathrm{E}(T_{n2}(j)) = o(\frac{1}{n^3|\boldsymbol{H}|^2}), \tag{3.48}$$

which also implies

$$\mathrm{E}(T_{n2}(i)T_{n1}(j)) - \mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n1}(j)) = o(\frac{1}{n^3|\boldsymbol{H}|^2}). \tag{3.49}$$

Since

$$\begin{aligned}
\mathrm{E}(T_{n2}(i)T_{n2}(j)) &= \frac{1}{n^2}E\left(\frac{K^2_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n^2(\boldsymbol{X}_j-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)f(\boldsymbol{X}_j)}\right) \\
&= \frac{1}{n^3|\boldsymbol{H}|^2}\left\{n|\boldsymbol{H}|\int\int K^2_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{u})\rho_n^2(\boldsymbol{v}-\boldsymbol{u})d\boldsymbol{u}d\boldsymbol{v}\right\} \\
&= \frac{K^2(\mathbf{0})\ \rho_{II}\ A(\Omega)}{n^3|\boldsymbol{H}|^2} + o(\frac{1}{n^3|\boldsymbol{H}|^2})\ \text{(by Proposition 3.1)},
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n2}(j)) \\
&= \frac{1}{n^2}E\left(\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\right) E\left(\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_j-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_j-\boldsymbol{X}_i)}{f(\boldsymbol{X}_j)}\right) \\
&= \frac{1}{n^3|\boldsymbol{H}|^2}\left\{n|\boldsymbol{H}|\int\int K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{u})\rho_n^2(\boldsymbol{v}-\boldsymbol{u})f(\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}\right\}^2 \\
&= \frac{K^2(\mathbf{0})\rho_I^2}{n^3|\boldsymbol{H}|^2} + o(\frac{1}{n^3|\boldsymbol{H}|^2})\ \text{(by Proposition 3.1)},
\end{aligned}$$

we get

$$\mathrm{E}(T_{n2}(i)T_{n2}(j)) - \mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n2}(j)) = O(\frac{1}{n^3|\boldsymbol{H}|^2}). \tag{3.50}$$

Similarly, by using the first two equations of Proposition 3.1, we can show

$$\mathrm{E}(T_{n2}(i)T_{n3}(j)) - \mathrm{E}(T_{n2}(i))\mathrm{E}(T_{n3}(j)) = O(\frac{1}{n^3|\boldsymbol{H}|^2}), \tag{3.51}$$

and

$$\mathrm{E}(T_{n3}(i)T_{n2}(j)) - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n2}(j)) = O(\frac{1}{n^3|\boldsymbol{H}|^2}). \qquad (3.52)$$

Finally,

$$\begin{aligned}
&\mathrm{E}(T_{n3}(i)T_{n3}(j)) - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n3}(j)) \\
= \; & \frac{1}{n^2}\mathrm{E}\left(\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\sum_{t\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_t-\boldsymbol{X}_j)\rho_n(\boldsymbol{X}_t-\boldsymbol{X}_i)}{f(\boldsymbol{X}_j)}\right) \\
& -\frac{1}{n^2}\mathrm{E}\left(\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\right)\mathrm{E}\left(\sum_{t\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_t-\boldsymbol{X}_j)\rho_n(\boldsymbol{X}_t-\boldsymbol{X}_i)}{f(\boldsymbol{X}_j)}\right) \\
= \; & \frac{1}{n^2}\mathrm{E}\left(\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_j)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_j)}{f(\boldsymbol{X}_i)f(\boldsymbol{X}_j)}\right) \\
& +\frac{1}{n^2}\mathrm{E}\left(\sum_{s\neq i,j}\sum_{t\neq s,i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)K_{\boldsymbol{H}}(\boldsymbol{X}_t-\boldsymbol{X}_j)\rho_n(\boldsymbol{X}_t-\boldsymbol{X}_j)}{f(\boldsymbol{X}_i)f(\boldsymbol{X}_j)}\right) \\
& -\frac{1}{n^2}\mathrm{E}\left(\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\right)\mathrm{E}\left(\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_j)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_j)}{f(\boldsymbol{X}_j)}\right) \\
& -\frac{1}{n^2}\mathrm{E}\left(\sum_{s\neq i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_s-\boldsymbol{X}_i)\rho_n(\boldsymbol{X}_s-\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\right)\mathrm{E}\left(\sum_{t\neq s,i,j}\frac{K_{\boldsymbol{H}}(\boldsymbol{X}_t-\boldsymbol{X}_j)\rho_n(\boldsymbol{X}_t-\boldsymbol{X}_j)}{f(\boldsymbol{X}_j)}\right).
\end{aligned}$$

Note that in the above, the second term cancels the fourth term. By applying the last two equations of Proposition 3.1, we have

$$\begin{aligned}
&\mathrm{E}(T_{n3}(i)T_{n3}(j)) - \mathrm{E}(T_{n3}(i))\mathrm{E}(T_{n3}(j)) \\
= \; & \frac{n-2}{n^3|\boldsymbol{H}|^2}(n^2|\boldsymbol{H}|^2\int\int\int K_{\boldsymbol{H}}(\boldsymbol{w}-\boldsymbol{u})\rho_n(\boldsymbol{w}-\boldsymbol{u})K_{\boldsymbol{H}}(\boldsymbol{w}-\boldsymbol{v})\rho_n(\boldsymbol{w}-\boldsymbol{v}) \\
& \cdot f(\boldsymbol{w})d\boldsymbol{u}d\boldsymbol{v}d\boldsymbol{w}) \\
& -\frac{n-2}{n^4|\boldsymbol{H}|^2}\left\{n|\boldsymbol{H}|\int\int K_{\boldsymbol{H}}(\boldsymbol{v}-\boldsymbol{u})\rho_n(\boldsymbol{v}-\boldsymbol{u})f(\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v}\right\}^2 \\
= \; & o(\frac{1}{n^3|\boldsymbol{H}|^2}). \qquad (3.53)
\end{aligned}$$

By substituting the results from (3.45) through (3.53) into (3.44), we get

$$\mathrm{Cov}(T_n(\boldsymbol{X}_i),T_n(\boldsymbol{X}_j)) = O(\frac{1}{n^3|\boldsymbol{H}|^2}) \text{ (for } i\neq j). \qquad (3.54)$$

By (3.41),

$$E(\frac{1}{n}\sum_{i=1}^{n}T_n(\boldsymbol{X}_i)) = \frac{K(\boldsymbol{0})(A(\Omega)+\rho_I)}{n|\boldsymbol{H}|} + o_p\left(\frac{1}{n|\boldsymbol{H}|}\right). \qquad (3.55)$$

By (3.43) and (3.54),

$$\begin{aligned}\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} T_n(\boldsymbol{X}_i)) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(T_n(\boldsymbol{X}_i)) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathrm{Cov}(T_n(\boldsymbol{X}_i), T_n(\boldsymbol{X}_j)) \\ &= \frac{1}{n}O(\frac{1}{n^2|\boldsymbol{H}|^2}) + \frac{n^2-n}{n^2}O(\frac{1}{n^3|\boldsymbol{H}|^2}) \\ &= O(\frac{1}{n^3|\boldsymbol{H}|^2}) = o(\frac{1}{n^2|\boldsymbol{H}|^2}).\end{aligned}$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} T_n(\boldsymbol{X}_i) = \frac{K(\mathbf{0})(A(\Omega)+\rho_I)}{n|\boldsymbol{H}|} + o_p(\frac{1}{n|\boldsymbol{H}|}).$$

From (3.38), The leading term of $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_{\boldsymbol{X}_i}^T \boldsymbol{\rho}_i$ is $\frac{1}{n}\sum_{i=1}^{n} T_n(\boldsymbol{X}_i)$. Hence Lemma 3.1 follows.

■

4 SELECTION OF BANDWIDTH WITH ESTIMATED COVARIANCES

In Chapter 3, we proposed some criteria for bandwidth selection. To apply these criteria, the covariances between errors must be completely known. In practice, the covariances have to be estimated from data. This will be the major issue in this chapter. In this chapter, we will assume a parametric model for the covariances so that estimation of the covariances reduces to estimation of the unknown parameters in the covariances model. The organization of this chapter is as follows. In section 4.1, we introduce some basic concepts for spatial random processes, including the semi-variogram, a partial linear semi-variogram model, and parametric semi-variogram model fitting. In section 4.2, we consider spatial data with multiple uncorrelated realizations, where at each design point there are repeated measurements of the response variable. In section 4.3, we consider spatial data collected on a grid.

4.1 Basic concepts

Suppose the random error process $\left\{\varepsilon(\boldsymbol{x}),\ \boldsymbol{x} \in I\!R^d\right\}$ is stationary and $\boldsymbol{x}$ is fixed. The process $\{\varepsilon(\boldsymbol{x})\}$ is called stationary if $\varepsilon(\boldsymbol{x})$ has zero mean, and the covariance between $\varepsilon(\boldsymbol{x})$ and $\varepsilon(\boldsymbol{x}+\boldsymbol{d})$ is a function of $\boldsymbol{d}$, the distance between $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{d}$. This implies that the variance of $\varepsilon(\boldsymbol{x})$ is the same for any $\boldsymbol{x}$. The function

$$c(\boldsymbol{d}) = \mathrm{Cov}(\varepsilon(\boldsymbol{x}+\boldsymbol{d}), \varepsilon(\boldsymbol{x})) \tag{4.1}$$

is called the covariogram of the error process.

In spatial statistics, the semi-variograiom is also an important concept. It is defined for intrinsically stationary processes, a wider class than stationary processes.

A random error process $\varepsilon(\boldsymbol{x})$ with zero mean is called intrinsically stationary if the variance of $\varepsilon(\boldsymbol{x}+\boldsymbol{d})-\varepsilon(\boldsymbol{x})$ is a function of $\boldsymbol{d}$ only (Cressie (1991, page 40)). The function

$$\gamma(\boldsymbol{d}) = \frac{1}{2}\mathrm{Var}(\varepsilon(\boldsymbol{x}+\boldsymbol{d}) - \varepsilon(\boldsymbol{x})) \tag{4.2}$$

is called the semi-variograiom of the error process. A stationary process is an intrinsically stationary process with homogeneous variance.

We will concentrate on the stationary processes. For a stationary process, there exists a simple relation between the semi-variogram and the covariogram:

$$c(\boldsymbol{d}) = \sigma^2 - \gamma(\boldsymbol{d}), \tag{4.3}$$

where $\sigma^2 = c(0)$ is the variance of $\varepsilon(\boldsymbol{x})$. In addition, we assume that the semi-variogram is isotropic. That is,

$$\gamma(\boldsymbol{d}) = \gamma_0(\|\boldsymbol{d}\|), \tag{4.4}$$

where $\gamma_0(\cdot)$ is a univariate function. In other words, $\gamma(\boldsymbol{d})$ is a function of the norm of $\boldsymbol{d}$, but it does not depend on the direction of $\boldsymbol{d}$. From (4.3), if the semi-variogram of a process is isotropic, then its covariogram is also a function of $\|\boldsymbol{d}\|$ only. Let $c(\boldsymbol{d}) = c_0(\|\boldsymbol{d}\|)$, then $c_0(\|\boldsymbol{d}\|) = \sigma^2 - \gamma_0(\|\boldsymbol{d}\|)$.

In order for a semi-variogram (or covariogram) function to be valid, it must satisfy the following condition. Given any k design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, and any k real numbers $a_1, \ldots, a_k$,

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_i c_0(\|\boldsymbol{x}_i - \boldsymbol{x}_i\|)a_j \geq 0.$$

The following are examples of some basic parametric isotropic semi-variogram models (see Cressie (1991, page 61)), with $\boldsymbol{\theta}$ as a vector of unknown parameters (called spatial dependence parameters). They are widely used in spatial statistics.

- *Exponential model*

$$\gamma(\boldsymbol{d};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{d} = \mathbf{0}, \\ a_0 + a_1\left(1 - \exp(-\alpha\|\boldsymbol{d}\|)\right), & \boldsymbol{d} \neq \mathbf{0}, \end{cases}$$

with $\boldsymbol{\theta} = (a_0, a_1, \alpha)'$, and $a_0 \geq 0$, $a_1 \geq 0$, $\alpha \geq 0$.

- *Spherical model*

$$\gamma(\boldsymbol{d};\boldsymbol{\theta}) = \begin{cases} 0, & \|\boldsymbol{d}\| = 0, \\ a_0 + a_1\left\{\frac{3}{2}\left(\frac{\|\boldsymbol{d}\|}{\alpha}\right) - \frac{1}{2}\left(\frac{\|\boldsymbol{d}\|}{\alpha}\right)^3\right\}, & 0 < \|\boldsymbol{d}\| \leq \alpha, \\ a_0 + a_1, & \|\boldsymbol{d}\| \geq \alpha, \end{cases}$$

with $\boldsymbol{\theta} = (a_0, a_1, \alpha)'$, and $a_0 \geq 0$, $a_1 \geq 0$, and $\alpha \geq 0$.

- *Rational quadratic model*

$$\gamma(\boldsymbol{d};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{d} = \mathbf{0}, \\ a_0 + a_1\left\{\frac{\|\boldsymbol{d}\|^2/\alpha}{1+\|\boldsymbol{d}\|^2/\alpha}\right\}, & \boldsymbol{d} \neq \mathbf{0}, \end{cases}$$

with $\boldsymbol{\theta} = (a_0, a_1, \alpha)'$, and $a_0 \geq 0$, $a_1 \geq 0$, and $\alpha \geq 0$.

These variogram models belong to a class of models that may be called "partial linear models". They can be written as

$$\gamma(\boldsymbol{d};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{d} = \mathbf{0}, \\ a_0 + a_1 g(\|\boldsymbol{d}\|, \alpha), & \boldsymbol{d} \neq \mathbf{0}, \end{cases} \tag{4.5}$$

with $\boldsymbol{\theta} = (a_0, a_1, \alpha)'$, and $a_0 \geq 0$, $a_1 \geq 0$, and $\alpha \geq 0$. Here $g(\cdot, \alpha)$ is continuous, $g(0, \alpha) = 0$, and $\lim_{\|\boldsymbol{d}\| \to \infty} g(\|\boldsymbol{d}\|, \alpha) = 1$.

In practice, it is often reasonable to assume $c(\infty, \boldsymbol{\theta}) = 0$, which means that data that are far apart are essentially uncorrelated. Hence $\gamma(\infty, \boldsymbol{\theta}) = \sigma^2$, and

$$a_0 + a_1 = \sigma^2.$$

From (4.3) and (4.5), it follows that the covariogram for partial linear models is

$$c(\boldsymbol{d},\boldsymbol{\theta}) = \begin{cases} a_0 + a_1, & \boldsymbol{d} = \boldsymbol{0}, \\ a_1\left(1 - g(\|\boldsymbol{d}\|, \alpha)\right), & \boldsymbol{d} \neq \boldsymbol{0}. \end{cases} \tag{4.6}$$

When $a_0 \neq 0$, the covariogram is not continuous at $\boldsymbol{d} = \boldsymbol{0}$, and a_0 is called the nugget effect.

At this point, let us assume that the errors $\varepsilon_i = \varepsilon(\boldsymbol{x}_i)$ are observable. Then estimation of the semi-variogram can be carried out in the following two steps.

The first step is to calculate the empirical semi-variogram. Given a distance d (a scalar), we define a set of index pairs:

$$S(d,t) = \{(i,i') : d - t \leq \|\boldsymbol{x}_i - \boldsymbol{x}_j\| < d + t\}, \tag{4.7}$$

where $t > 0$ is the tolerance value, usually a small number. If the distance between a pair of design points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is in the tolerance interval $(d-t, d+t)$, then their index pair is in $S(d,t)$. Let $n(d,t)$ be the number of elements in the set $S(d,t)$. Then the empirical semi-variogram is

$$\hat{\gamma}_0(d) = \frac{1}{2n(d,t)} \sum_{(i,j)\in S(d,t)} (\varepsilon_i - \varepsilon_j)^2. \tag{4.8}$$

Notice that for $(i,j) \in S(d,t)$, if $\gamma_0(\cdot)$ is continuously differentiable,

$$\mathrm{E}\left\{\frac{1}{2}(\varepsilon_i - \varepsilon_j)^2\right\} = \frac{1}{2}\mathrm{Var}(\varepsilon_i - \varepsilon_j) = \gamma_0(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) = \gamma_0(d) + O(t).$$

So if the tolerance value t is small enough, the empirical semi-variogram $\hat{\gamma}_0(d)$ is an approximately unbiased estimator for the semi-variogram $\gamma_0(d)$. The bias is related to the value of t and the first derivative of the semi-variogram function $\gamma_0(\cdot)$ near the distance d. Following Cressie (1991, page 99), we select a sequence of distances: $0 < d_1 < d_2 < \cdots < d_K$. For every d_k $(k = 1, \ldots, K)$, calculate the empirical semi-variogram $\hat{\gamma}_0(d_k)$. When selecting the sequence $\{d_k\}$, d_1 should be close to 0, so that the nugget effect of the semi-variogram can be estimated with satisfactory precision; Also, the biggest distance d_K is usually less than a half of the maximum distance between any pair of design points.

The second step is to fit a parametric function $\gamma_0(d, \boldsymbol{\theta})$ to the empirical semi-variogram. The estimator of the spatial dependence parameter $\boldsymbol{\theta}$ is obtained by finding the value of $\boldsymbol{\theta}$ that minimizes

$$\sum_{k=1}^{K} \frac{n(d_k, t)}{\gamma_0^2(d_k; \boldsymbol{\theta})} \left(\hat{\gamma}_0(d_k) - \gamma_0(d_k; \boldsymbol{\theta})\right)^2 . \tag{4.9}$$

Now, let us consider spatial data that consist of a spatial trend (the mean function) and spatially correlated errors. Since the errors are not observable, the above procedure for semi-variogram estimation can not be applied directly, unless there is no spatial trend. When a spatial trend is present, the empirical semi-variogram may be calculated from residuals. If we assume a parametric form for the spatial trend, then the technique of universal kriging can be used. In universal kriging, an iterative weighted least squares procedure can be used to estimate the spatial trend and the semi-variogram of the errors. The basic idea is as follows. Starting from an ordinary least squares estimates for parameters in the spatial trend model, obtain the residuals and use them for estimating the semi-variogram by minimizing (4.9). Use the estimated spatial correlations to update the estimates of the spatial trend parameters using generalized least squares estimation. Then, obtain new residuals. This procedure is iterated until convergence is reached for the estimates of the parameters in the spatial trend model.

In a nonparametric approach, however, no parametric form is specified for the spatial trend. The nonparametric analogue of parametric ordinary least squares estimation may be considered as cross-validation (in uncorrected version), since in cross-validation, the correlations between errors are ignored as well, just as in parametric ordinary least squares estimation. Unfortunately, as shown in chapter 3 (both theoretically and numerically), cross-validation tends to select bandwidths that are too small, resulting in a rough estimate of the spatial trend when data are positively correlated. The residuals from a cross-validation bandwidth are usually a poor substitute for the true random errors. Thus starting from the residuals of cross-validation often yields poor results.

One approach for selecting a pilot bandwidth, analogous to an approach in univariate cases suggested by some authors, might be as follows. Select a pilot bandwidth which results in a "smooth" and "good" fit of the original data. Then use the residuals from this pilot fit to estimate the spatial dependence parameters by semi-variogram fitting. However, there are lots of bandwidths which may give a "smooth" and "good" fit of the original data, and it is up to the user to subjectively choose a particular one. So this approach is not easily evaluated.

One case where consistent estimation of spatial correlations can be done by fitting an empirical variogram is spatial data with repeated measurements. This will be discussed in the next section.

4.2 Spatial data with repeated measurements

4.2.1 Estimating spatial dependence parameters by fitting a parametric model to the empirical semivariograms

The estimation of growth curves has been studied extensively in parametric situations. Hart and Wehrly (1986) applied a noparametric method, namely, the Gasser-Müller estimator, to this important problem. In this section, we consider a similar situation for spatial data with repeated measurements. This kind of setting might come from image processing, where we want to uncover the true image of a target from its several independent snapshots, which might have been contaminated by correlated noises.

We suppose that at every design point $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$, repeated measurements $Y_{i,1}, \ldots, Y_{i,J}$ for the response are available. Let

$$Y_{i,j} = m(\boldsymbol{x}_i) + \varepsilon_{i,j} \;\; (i = 1, \ldots, n;\; j = 1, \ldots, J), \tag{4.10}$$

where $\varepsilon_{i,j}$ have zero mean and satisfy

$$\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = \begin{cases} 0, & j \neq j', \\ c_0(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|), & j = j', \end{cases} \tag{4.11}$$

and $m(\cdot)$ is the fixed spatial trend. Here j ($j = 1, \ldots, J$) is the index corresponding to one complete realization of the process $Y(\boldsymbol{x})$. For a fixed j, $\{Y_{i,j}, (i = 1 \ldots, n)\}$ represents one complete realization. Equation (4.11) implies that different realizations are not correlated to each other, and that all the realizations have the same covariance structure.

Consider the mean of the realizations at the i-th design point, $\overline{Y}_{i\cdot} = \sum_{j=1}^{J} Y_{i,j}/J$. From equation (4.10), we have

$$\overline{Y}_{i\cdot} = m(\boldsymbol{x}_i) + \bar{\varepsilon}_{i\cdot}, \tag{4.12}$$

with

$$\mathrm{Cov}(\overline{Y}_{i\cdot}, \overline{Y}_{i'\cdot}) = \mathrm{Cov}(\bar{\varepsilon}_{i\cdot}, \bar{\varepsilon}_{i'\cdot}) = \frac{1}{J} c_0(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|). \tag{4.13}$$

Now the problem of estimating $m(\cdot)$ becomes a problem of fitting a smooth surface through the sample means of the realizations, i.e. a new data set $\left\{(\boldsymbol{x}_i, \overline{Y}_{i\cdot}),\ \ i = 1, \ldots, n\right\}$. If we can estimate the function $c_0(\cdot)$, the bandwidth selection criteria proposed in chapter 3 can be directly applied here for estimating $m(\cdot)$.

It can be easily seen that

$$\begin{aligned} \mathrm{E}\left\{\frac{1}{J-1}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i\cdot})(Y_{i'j} - \overline{Y}_{i'\cdot})\right\} &= \frac{1}{J-1}\sum_{j=1}^{J} \mathrm{E}\left\{\varepsilon_{i,j} - \bar{\varepsilon}_{i\cdot})(\varepsilon_{i'j} - \bar{\varepsilon}_{i'\cdot})\right\} \\ &= \frac{1}{J-1}\sum_{j=1}^{J} \mathrm{Cov}(\varepsilon_{i,j} - \bar{\varepsilon}_{i\cdot}, \varepsilon_{i'j} - \bar{\varepsilon}_{i'\cdot}) \\ &= c_0(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|), \end{aligned}$$

and

$$\mathrm{E}\left\{\frac{1}{n(J-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i\cdot})^2\right\} = \mathrm{E}\left\{\frac{1}{n(J-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(\varepsilon_{i,j} - \bar{\varepsilon}_{i\cdot})^2\right\} = \sigma^2.$$

Hence

$$\frac{1}{n(J-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i\cdot})^2-\frac{1}{J-1}\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i\cdot})(Y_{i'j}-\overline{Y}_{i'\cdot}) \tag{4.14}$$

is an unbiased estimator for the semi-variogram $\gamma_0(\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|)$. Therefore, the empirical semi-variogram $\hat{\gamma}_0(d)$ can be defined as

$$\begin{aligned}
&\hat{\gamma}_0(d)\\
=\;&\frac{1}{n(d,t)}\sum_{(i,i')\in S(d,t)}\left\{\frac{1}{n(J-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i\cdot})^2-\frac{1}{J-1}\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i\cdot})(Y_{i'j}-\overline{Y}_{i'\cdot})\right\}\\
=\;&\frac{1}{n(J-1)}\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i\cdot})^2-\frac{1}{n(d,t)}\sum_{(i,i')\in S(d,t)}\left\{\frac{1}{J-1}\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i\cdot})(Y_{i'j}-\overline{Y}_{i'\cdot})\right\}.
\end{aligned} \tag{4.15}$$

where the set of indices $S(d,t)$ is defined in (4.7), and $n(d,t)$ is the number of elements in $S(d,t)$. Suppose we know that the semi-variogram of the errors has a parametric form given by $\gamma_0(d,\boldsymbol{\theta})$, then a weighted least squares estimator $\hat{\boldsymbol{\theta}}$ for the spatial dependence parameter $\boldsymbol{\theta}$ can be obtained by minimizing (4.9) of section 4.1. From (4.13), we obtain the estimator of the covariance between $\overline{Y}_{i\cdot}$ and $\overline{Y}_{i'\cdot}$: $\frac{1}{J}(\hat{\sigma}^2-\gamma_0(\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|,\hat{\boldsymbol{\theta}}))$. With the estimators of the covariances, the bandwidth selection criteria proposed in Chapter 3 can then be applied to the data set $\{(\boldsymbol{x}_i,\overline{Y}_{i\cdot}),\ i=1,\ldots,n\}$.

4.2.2 Simulation study

First, we use a simulated example to illustrate the above method. We randomly generate 400 design points $\boldsymbol{x}_i=(x_{i,1},x_{i,2})$ in the rectangle: $0<x_1<1$, $0<x_2<1$ according to a uniform distribution. Then, we generate 2 repeated measurements of the response variable at these points by

$$Y_{i,j}=\sin(2\pi x_{i,1})+4(x_{i,2}-0.5)^2+\varepsilon_{i,j},\quad (i=1,\ldots,400;\ j=1,2), \tag{4.16}$$

where the errors $\varepsilon_{i,j}$ have zero mean and satisfy

$$\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = \begin{cases} 0, & j \neq j', \\ \sigma^2 \alpha^{20\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|}, & j = j'. \end{cases} \tag{4.17}$$

with $\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| = \sqrt{(x_{i,1} - x_{i',1})^2 + (x_{i,2} - x_{i',2})^2}$.

In this example, we take $\sigma = 0.5$ and $\alpha = 0.7$ to generate $\{Y_{i,j}\}$. One pair of complete realizations are plotted in figure 4.1(a) and figure 4.1(b). The average of these two realizations is plotted in figure 4.1(c). Estimation of the mean function is now a problem of smoothing the data in figure 4.1(c). The empirical semi-variogram is then calculated according to equation (4.15) at distances $d_k = 0.02k - 0.01$ $(k = 1, \ldots, 30)$, with the tolerance value $t = 0.005$. Suppose we know the parametric form of the covariance function of the errors, then we know the parametric form of the semi-variogram

$$\gamma_0(d, \boldsymbol{\theta}) = \sigma^2 - \sigma^2 \alpha^{20d},$$

where $\boldsymbol{\theta} = (\sigma, \alpha)^T$ is unknown and needs to be estimated from data. By fitting the empirical semi-variogram with this parametric semi-variogram model, we obtain estimates of the spatial dependence parameters: $\hat{\sigma} = 0.51$, $\hat{\alpha} = 0.72$. The empirical semi-variogram and the fitted curve are plotted in figure 4.1(d).

Next, local linear regression with the Epanechnikov kernel is used to estimate the true mean function. For the moment, candidates for the bandwidth matrix are taken to be spherical: $\boldsymbol{H} = hI_2$, with h taking values 0.1, 0.2, 0.3, 0.4, and 0.5. The estimates $\hat{\sigma}$ and $\hat{\alpha}$ are used in calculation of the GCV and the corrected GCV criteria. The values of h chosen by the GCV and the corrected GCV criteria are 0.1 and 0.2, resulting in the average squared errors of 0.0883 and 0.0710, respectively. In this particular example, the corrected GCV criterion outperforms the GCV criterion. Figure 4.2 displays the estimates of the mean function provided by these two criteria and the true surface. It can be seen that the corrected GCV criterion provides an estimate that is smoother, and closer to the true surface than the GCV criterion.

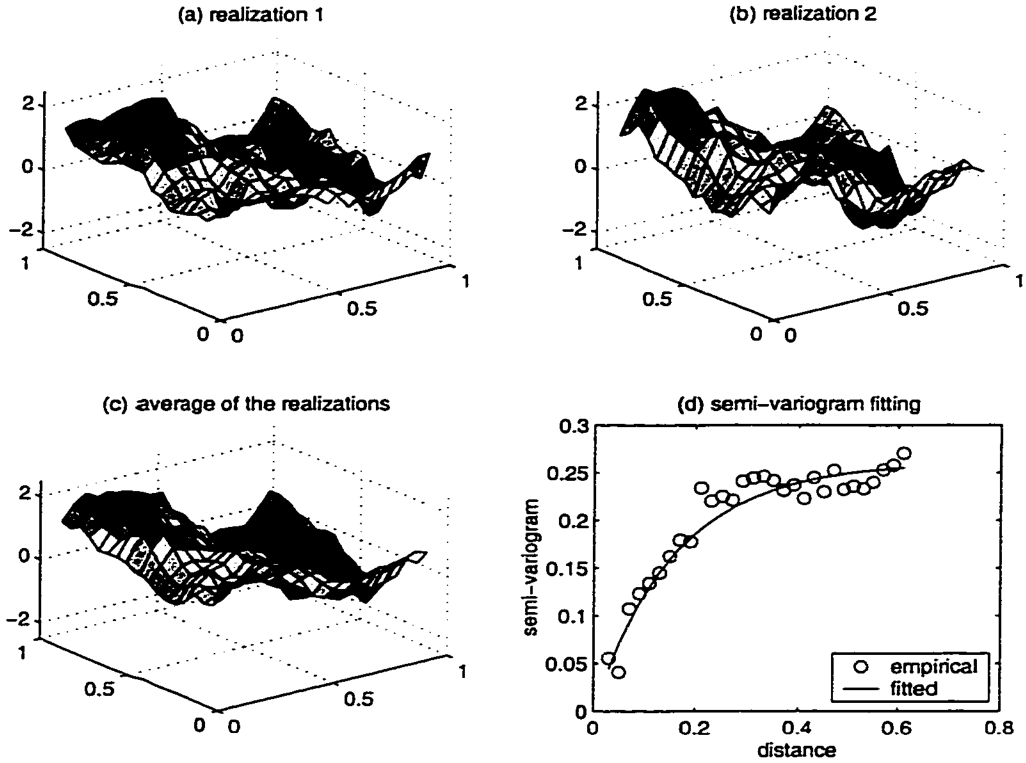

Figure 4.1 A simulated example with 2 repeated measurements

In the following, a larger simulation study is used to numerically show the benefit of the proposed method in handling spatial data with repeated measurements. The GCV criterion, the corrected C_L criterion, and the corrected GCV criterion will be used to select the bandwidth parameters. Two kinds of bandwidth searching schemes will be considered: searching from diagonal bandwidths $\boldsymbol{H} = \text{diag}(h_1, h_2)$, and searching from spherical bandwidths $\boldsymbol{H} = h\boldsymbol{I}_2$ only. In this study, h_1, h_2, and h take values from 0.1, 0.2, 0.3, 0.4, and 0.5. In addition to the case when the parametric covariance function is exactly known, we will investigate the case when the parametric covariance function is misspecified as well.

First, let us assume that the parametric form of the covariance function used to generate data, (4.17), is known but the parameters σ and α are not known. The setting

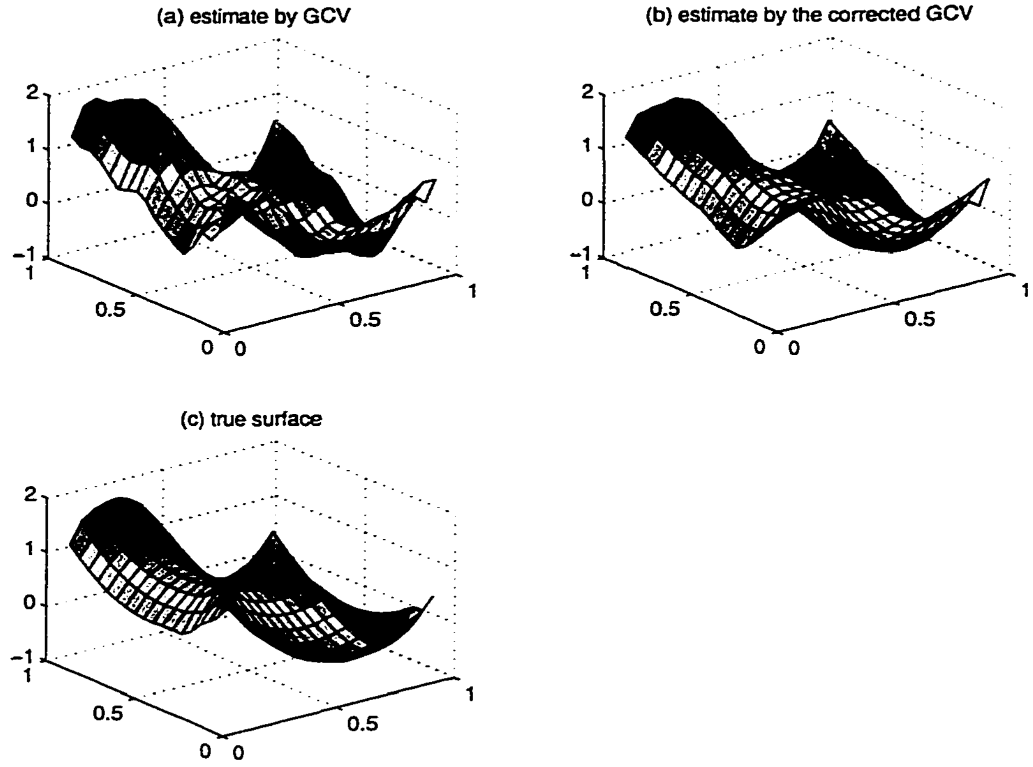

Figure 4.2 Estimates of the surface using local linear regression

is the same as in the previous example, except that different values of σ and α will be considered. In this simulation study, σ is either 0.3 or 0.5 and α is set at either 0, 0.3, or 0.7. So we have 6 different cases based on the combinations of the values of these two parameters. For each case, 50 pairs of complete replications are simulated (a new set of design points is generated for each of the 50 replications). In addition, for each of the two cases: $(\sigma, \alpha) = (0.5, 0)$ and $(\sigma, \alpha) = (0.5, 0.3)$, we simulate 50 more replications to show that 50 replications are enough to obtain stable simulation results. We will see that the second 50 replications for each of these two cases have similar results with the first 50 replications in estimates of the spatial dependence parameters, selected diagonal bandwidths, and the means of average squared errors. All these simulation results are summarized in tables 4.1 to 4.5.

Table 4.1 shows the means and standard deviations of the estimates of σ and α for the 50 simulations. The weighted least squares estimation procedure that minimizes (4.9) appears to provide reasonable estimates even though only 2 complete realizations are simulated. In this case, variation for estimates of σ and α is larger when α is larger, i.e., when there are stronger positive spatial correlations.

Table 4.1 The means and the standard deviations (in parentheses) of the estimates for σ and α by semi-variogram fitting (the numbers in italics are results of the second 50 replications).

σ	α	$\hat{\sigma}$	$\hat{\alpha}$
0.3	0	0.3013 (0.0122)	0.0079 (0.0155)
0.3	0.3	0.2972 (0.0163)	0.2874 (0.0859)
0.3	0.7	0.3241 (0.0328)	0.7011 (0.1166)
0.5	0	0.5005 (0.0200)	0.0064 (0.0137)
0.5	0.3	0.5012 (0.0235)	0.3181 (0.0743)
0.5	0.7	0.5070 (0.0770)	0.6948 (0.1028)
0.5	*0*	*0.5001 (0.0190)*	*0.0100 (0.0149)*
0.5	*0.3*	*0.4959 (0.0263)*	*0.3018 (0.0619)*

Table 4.2 contains the average values of bandwidth parameters h_1 and h_2 selected by the GCV criterion, the corrected C_L criterion, and the corrected GCV criterion. For purpose of comparison, the average values of the optimal h_1 and h_2 are listed in the table as well. They are selected from possible bandwidth candidates that minimize (3.24) (using the true values of σ, α, and the true mean function). From this table, we see that when data are positively correlated ($\alpha > 0$), the GCV criterion tends to choose $h_1 = 0.1$ and $h_2 = 0.1$, which correspond to the smallest bandwidth region. The average values of h_1 and h_2 chosen by the corrected C_L criterion and the corrected GCV criterion are very similar, and much closer to the optimal bandwidth than the GCV selections when $\alpha > 0$. When data are not correlated, the 3 criteria provide similar results.

Table 4.3 gives the average values of bandwidth parameter h using spherical bandwidth search only. Conclusions drawn from this table are similar to those presented above.

Table 4.2 The average values of (h_1, h_2) selected by 3 criteria, and the average values of the optimal (h_1, h_2) by searching diagonal bandwidths (the numbers in italics are results of the second 50 replications).

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	(0.1000, 0.2080)	(0.1000, 0.2260)	(0.1000, 0.2340)	(0.1000, 0.2000)
0.3	0.3	(0.1000, 0.1000)	(0.1040, 0.2480)	(0.1040, 0.2420)	(0.1000, 0.3000)
0.3	0.7	(0.1000, 0.1000)	(0.1000, 0.2760)	(0.1040, 0.2840)	(0.1000, 0.3000)
0.5	0	(0.1020, 0.2840)	(0.1160, 0.2920)	(0.1180, 0.2900)	(0.1000, 0.3000)
0.5	0.3	(0.1000, 0.1020)	(0.1660, 0.3200)	(0.1700, 0.3220)	(0.2000, 0.3000)
0.5	0.7	(0.1000, 0.1000)	(0.1680, 0.3440)	(0.1700, 0.3520)	(0.2000, 0.3040)
0.5	*0*	*(0.1060, 0.2660)*	*(0.1320, 0.2840)*	*(0.1280, 0.2880)*	*(0.1000, 0.3000)*
0.5	*0.3*	*(0.1000, 0.1000)*	*(0.1580, 0.2820)*	*(0.1700, 0.2900)*	*(0.2000, 0.3000)*

Table 4.3 The average values of h selected by 3 criteria, and the average values of the optimal h by searching spherical bandwidths only.

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.1260	0.1420	0.1680	0.1000
0.3	0.3	0.1000	0.1760	0.1780	0.2000
0.3	0.7	0.1000	0.1720	0.1940	0.2000
0.5	0	0.2000	0.1980	0.2000	0.2000
0.5	0.3	0.1000	0.2000	0.2040	0.2000
0.5	0.7	0.1000	0.2100	0.2260	0.2000

Table 4.4 shows the mean average squared errors of the local linear estimator using the diagonal bandwidths selected by the 3 different criteria, and using the optimal bandwidths. When data are correlated, the corrected C_L criterion and the corrected GCV criterion provide smaller mean average squared errors than the GCV criterion does, and the average squared errors by the corrected C_L criterion and the corrected GCV criterion are very close to the optimal bandwidths. When data are uncorrelated, the 3 criteria perform equally well.

Table 4.5 shows the mean average squared errors of the local linear estimator using the spherical bandwidths selected by the 3 different criteria, and using the optimal spherical bandwidths. By comparing this table with Table 4.4, if we fix the value of σ

and the value of α, for each of the 3 criteria, we see that the spherical bandwidths search gives slightly larger mean average squared errors.

Next, we consider the situation when the parametric covariance function is misspecified. We make a slight change in the simulation study. The data are generated in the same way as above, except that the covariance function used now is not (4.17) but the following spherical covariance function:

$$\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = \begin{cases} 0 & j \neq j' \\ \tilde{\sigma}^2 \left(1 - \frac{3}{2}\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|}{\tilde{\alpha}}\right) + \frac{1}{2}\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|}{\tilde{\alpha}}\right)^3\right) 1_{\{\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\| < \tilde{\alpha}\}} & j = j'. \end{cases} \tag{4.18}$$

When estimating the covariances between the errors, we use the misspecified model (4.17).

Table 4.4 The mean average squared errors using the bandwidths selected by 3 criteria, and using the optimal bandwidths by searching diagonal bandwidths (the numbers in italics are results of the second 50 replications).

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.0054	0.0054	0.0055	0.0052
0.3	0.3	0.0202	0.0142	0.0142	0.0135
0.3	0.7	0.0337	0.0276	0.0280	0.0264
0.5	0	0.0122	0.0125	0.0124	0.0118
0.5	0.3	0.0509	0.0290	0.0287	0.0267
0.5	0.7	0.0887	0.0708	0.0715	0.0662
0.5	*0*	*0.0125*	*0.0132*	*0.0130*	*0.0119*
0.5	*0.3*	*0.0510*	*0.0291*	*0.0286*	*0.0262*

Table 4.6 displays the average values of the diagonal bandwidth parameters h_1 and h_2 selected by the GCV criterion, the corrected C_L criterion, and the corrected GCV criterion, as well as the average values of the optimal h_1 and h_2. When calculating the corrected C_L criterion and the corrected GCV criterion, we use the estimates of covariances based on the misspecified model (4.17). When calculating the optimal h_1 and h_2, the true spatial dependence parameters $\tilde{\sigma}$ and $\tilde{\alpha}$, and the true covariance model

Table 4.5 The mean average squared errors using the bandwidths selected by 3 criteria, and using the optimal bandwidths by searching spherical bandwidths only.

σ	α	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.0090	0.0093	0.0096	0.0086
0.3	0.3	0.0202	0.0186	0.0181	0.0174
0.3	0.7	0.0337	0.0323	0.0315	0.0308
0.5	0	0.0142	0.0144	0.0142	0.0142
0.5	0.3	0.0512	0.0293	0.0300	0.0284
0.5	0.7	0.0887	0.0727	0.0752	0.0700

(4.18) are used. From this table, the average values of h_1 and h_2 chosen by the corrected C_L criterion and the corrected GCV criterion are very close, and much closer to the optimal values than the GCV selections when $\alpha > 0$. Also, when data are correlated, the GCV criterion tends to choose $h_1 = 0.1$ and $h_2 = 0.1$, which correspond to the smallest bandwidth region.

Table 4.6 The average values of (h_1, h_2) selected by 3 criteria, and the average values of the optimal (h_1,h_2) by searching diagonal bandwidths and with misspecified covariance function.

$\tilde{\sigma}$	$\tilde{\alpha}$	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	(0.1000, 0.2080)	(0.1000, 0.2260)	(0.1000, 0.2340)	(0.1000, 0.2000)
0.3	0.15	(0.1000, 0.1000)	(0.1000, 0.2260)	(0.1040, 0.2080)	(0.1000, 0.3000)
0.3	0.30	(0.1000, 0.1000)	(0.1020, 0.2660)	(0.1020, 0.2380)	(0.1000, 0.3000)
0.5	0	(0.1020, 0.2840)	(0.1160, 0.2920)	(0.1180, 0.2900)	(0.1000, 0.3000)
0.5	0.15	(0.1000, 0.1000)	(0.1400, 0.3300)	(0.1460, 0.3220)	(0.2000, 0.3000)
0.5	0.30	(0.1000, 0.1000)	(0.1620, 0.2860)	(0.1540, 0.2800)	(0.2000, 0.3700)

Table 4.7 displays the the average values of the spherical bandwidth parameter h. Table 4.8 shows the the mean average squared errors of the local linear estimator using the diagonal bandwidths selected by the 3 different criteria, and using the optimal bandwidths. From this table, the corrected C_L criterion and the corrected GCV criterion outperform the GCV criterion when data are correlated. They perform equally well as the GCV criterion when data are uncorrelated.

Table 4.7 The average values of h selected by 3 criteria, and the average values of the optimal h by searching spherical bandwidths only and with misspecified covariance function.

$\tilde{\sigma}$	$\tilde{\alpha}$	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.1260	0.1420	0.1680	0.1000
0.3	0.15	0.1000	0.1460	0.1240	0.2000
0.3	0.30	0.1000	0.1780	0.1600	0.2000
0.5	0	0.2000	0.1980	0.2000	0.2000
0.5	0.15	0.1000	0.1980	0.1880	0.2000
0.5	0.30	0.1000	0.1940	0.2160	0.2000

Table 4.8 The mean average squared errors using the bandwidths selected by 3 criteria, and using the optimal bandwidths by searching diagonal bandwidths and with misspecified covariance function).

$\tilde{\sigma}$	$\tilde{\alpha}$	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.0054	0.0054	0.0055	0.0052
0.3	0.15	0.0202	0.0142	0.0142	0.0135
0.3	0.30	0.0337	0.0276	0.0280	0.0264
0.5	0	0.0122	0.0125	0.0124	0.0118
0.5	0.15	0.0509	0.0290	0.0287	0.0267
0.5	0.30	0.0887	0.0708	0.0715	0.0662

Table 4.9 displays the mean average squared errors of the local linear estimator using the spherical bandwidth search only. The restricted search gives larger mean average squared errors for the local linear estimator.

In summary, even in the situation when the model of the covariances between errors is misspecified, the criteria proposed in Chapter 3 still work well.

4.3 Spatial data on a grid

4.3.1 Estimating spatial dependence parameters by differencing

When there is only one realization at each point in the design, estimation of spatial correlation is more difficult. Instead of estimating parameters in the covariance model using residuals from an initial estimate of the mean function (a pilot fit), several au-

Table 4.9 The mean average squared errors using the bandwidths selected by 3 criteria, and using the optimal bandwidths by searching spherical bandwidths only and with misspecified covariance function.

$\tilde{\sigma}$	$\tilde{\alpha}$	GCV	corrected C_L	corrected GCV	Optimal
0.3	0	0.0090	0.0093	0.0096	0.0086
0.3	0.15	0.0202	0.0186	0.0181	0.0174
0.3	0.30	0.0337	0.0323	0.0315	0.0308
0.5	0	0.0142	0.0144	0.0142	0.0142
0.5	0.15	0.0512	0.0293	0.0300	0.0284
0.5	0.30	0.0887	0.0727	0.0752	0.0700

thors have discussed the method of differencing in the case of univariate, equally spaced designs. Müller and Stadmüller (1988) considered the setting where the errors form a moving average process. Gasser *et al.* (1986) used differencing to estimate the variance function in a heteroscedastic regression model. Hart (1989, 1991) discussed more general cases where the errors form a zero mean, stationary time series.

Inspired by these researches, we will apply the differencing technique for spatial data collected on a grid. For simplicity, we focus on the 2-dimensional situation.

Suppose the design region Ω is a $n_1 \times n_2$ equally spaced grid in the rectangle:

$$\left\{\boldsymbol{x} \in I\!R^2 : \boldsymbol{x} = (x_1, x_2)^T,\ 0 \le x_1 \le a,\ \ 0 \le x_2 \le b\right\}.$$

Let $\delta_1 = a/n_1$, $\delta_2 = b/n_2$, and the design points $(x_{i,1}, x_{j,2})$ are $x_{i,1} = (i - 0.5)\delta_1$, $x_{j,2} = (j - 0.5)\delta_2$ $(i = 1, \ldots, n_1;\ j = 1, \ldots, n_2)$. So the total number of design points is $N = n_1 n_2$. We assume $n_k \to \infty$ $(k = 1, 2)$, and $n_2/n_1 \to r$ (a constant). We consider the model

$$Y_{i,j} = m_{i,j} + \varepsilon_{i,j} \quad (i = 1, \ldots, n_1;\ j = 1, \ldots, n_2), \tag{4.19}$$

where $m_{i,j} = m(x_{i,1}, x_{j,2})$ with $m(\cdot, \cdot)$ as a fixed spatial trend (or the mean function), and the errors $\varepsilon_{i,j}$ form a zero mean, stationary process with covariance:

$$\text{Cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = c_{n_1,n_2}(x_{i,1} - x_{i',1}, x_{j,2} - x_{j',2}) = \sigma^2 \rho_{n_1,n_2}(x_{i,1} - x_{i',1}, x_{j,2} - x_{j',2}).$$

In particular, we assume

$$\begin{aligned}\rho_{n_1,n_2}(x_{i,1} - x_{i',1},\ x_{j,2} - x_{j',2}) &= \rho\left(n_1 r_1(x_{i,1} - x_{i',1}),\ n_1 r_2(x_{j,2} - x_{j',2})\right) \\ &= \rho\left((i-i')r_1 a,\ (j-j')r_2 b\right),\end{aligned}$$

where ρ is a valid correlation function defined in $I\!R^2$, $r_1 > 0$ and $r_2 > 0$ are constants. Therefore, the correlation between $\varepsilon_{i,j}$ and $\varepsilon_{i',j'}$ is assumed to be a function of the difference between their indices. Let $\rho_\varepsilon(k,\ l) = \rho\,(kr_1 a,\ lr_2 b)$, then ρ_ε is a valid correlation function defined on the expanding grid of integers. The corresponding covariance function is $c_\varepsilon(k,\ l) = \sigma^2 \rho_\varepsilon(k,\ l)$. Thus the errors $\varepsilon_{i,j}$ can be treated as a stationary process on the expanding grid of integers. This setting is analogous to the univariate case described in Hart (1989, 1991). We assume that the correlation function $\rho_\varepsilon(k,\ l)$ goes to 0 as k or l goes to ∞.

Define the second order differences of the $\{Y_{i,j}\}$ as

$$\Delta_Y(i,j) = Y_{i,j} - \frac{1}{4}(Y_{i,j-1} + Y_{i,j+1} + Y_{i-1,j} + Y_{i+1,j}) \tag{4.20}$$

for $(i = 2, \ldots, n_1 - 1;\ j = 2, \ldots, n_2 - 1)$. Assume that the mean function m is twice continuously differentiable, and let

$$M_1 \equiv \max_\Omega \left(\left| \frac{\partial^2 m}{\partial x_1^2} \right| \right), \quad \text{and} \quad M_2 \equiv \max_\Omega \left(\left| \frac{\partial^2 m}{\partial x_2^2} \right| \right).$$

Define the second order difference of the mean function as

$$\Delta_m(i,j) = m_{i,j} - \frac{1}{4}(m_{i,j-1} + m_{i,j+1} + m_{i-1,j} + m_{i+1,j}).$$

By Taylor expansion, we can show

$$|\Delta_m(i,j)| \leq M_1 \delta_1^2 + M_2 \delta_2^2. \tag{4.21}$$

This inequality implies $\Delta_m(i,j) = O(1/N)$. Define the second order difference of the errors

$$\eta_{i,j} = \varepsilon_{i,j} - \frac{1}{4}(\varepsilon_{i,j-1} + \varepsilon_{i,j+1} + \varepsilon_{i-1,j} + \varepsilon_{i+1,j}). \tag{4.22}$$

Then

$$\Delta_Y(i,j) = \eta_{i,j} + \Delta_m(i,j) = \eta_{i,j} + O(1/N). \tag{4.23}$$

Besides the second order differencing, other differencing schemes are also possible. For example, we may want to consider the third order differences of the $\{Y_{i,j}\}$:

$$\begin{aligned} \Delta_Y^{(1)}(i,j) &= Y_{i,j} - \frac{1}{2}(Y_{i,j-1} + Y_{i,j+1} + Y_{i-1,j} + Y_{i+1,j}) \\ &\quad + \frac{1}{4}(Y_{i-1,j-1} + Y_{i-1,j+1} + Y_{i+1,j-1} + Y_{i+1,j+1}), \end{aligned} \tag{4.24}$$

for $(i = 2, \ldots, n_1 - 1;\ j = 2, \ldots, n_2 - 1)$. Note that a linear mean function can be completely removed by the second order differencing, and a quadratic mean function can be completely removed by the third order differencing. As the grid becomes finer (n_1 and n_2 become larger), differencing can effectively remove the spatial trend, so that the difference of data can well approximate the difference of the errors. Another example is that the spatial trend $m(x_1, x_2)$ is known to be an additive model, i.e.,

$$m(x_1, x_2) = m_1(x_1) + m_2(x_2).$$

In this case, if the difference is defined as

$$\Delta_Y^{(2)}(i,j) = Y_{i,j} + Y_{i+1,j+1} - Y_{i,j+1} - Y_{i+1,j}$$

for $(i = 1, \ldots, n_1 - 1;\ j = 1, \ldots, n_2 - 1)$, then the additive spatial trend $m(x_1, x_2)$ will be completely removed.

In the following discussion, we will concentrate on the second order difference defined by equation (4.23) only. Our problem is how to estimate the covariance of $\varepsilon_{i,j}$ based on the second order difference of data, $\Delta_Y(i,j)$. We assume the parametric form of the covariance of the errors is $c_\varepsilon(k, l, \boldsymbol{\theta})$, with unknown $\boldsymbol{\theta}$, the vector of spatial dependence parameters. Then the problem becomes how to estimate $\boldsymbol{\theta}$. Our method will be related to concepts and techniques in frequency domain estimation.

First, similar to Priestley (1994, page 720), we define the two dimensional spectral density for the zero mean, stationary process $\{\varepsilon_{i,j}\}$, which is the discrete Fourier transformation of its covariance function:

$$S_\varepsilon(\omega_1, \omega_2, \boldsymbol{\theta}) = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} c_\varepsilon(u, v, \boldsymbol{\theta}) \exp(-i(u\omega_1 + v\omega_2)), \tag{4.25}$$

with $\omega_1, \omega_2 \in [-\pi, \pi]$. The spectral density in Priestley (1994, page 720) is the spectral density here divided by the constant $(2\pi)^2$.

Note that $\eta_{i,j}$, as the second order difference of the errors, is also stationary, so that its power spectral density is well defined as well. From (4.22), we get

$$S_\eta(\omega_1, \omega_2, \boldsymbol{\theta}) = \left(1 - \frac{1}{2}(\cos(\omega_1) + \cos(\omega_2))\right)^2 S_\varepsilon(\omega_1, \omega_2, \boldsymbol{\theta}). \tag{4.26}$$

At this point, let us pretend that $\{\eta_{i,j}\}$ is observable. Since $\{\eta_{i,j}\}$ is stationary, some known results for stationary processes on a grid can be applied here directly.

We consider how to estimate the sample covariance of $\{\eta_{i,j}\}$ at lag (u, v). One possible estimator is

$$\hat{c}_\eta(u, v) = \frac{1}{(n_1 - 2)(n_2 - 2)} \sum_{i=2}^{n_1 - 1 - u} \sum_{j=2}^{n_2 - 1 - v} \eta_{i,j}\eta_{i+u,j+v}.$$

This estimator has a bias of order $O(N^{-1/2})$ due to the boundary effect. As pointed out by several authors, boundary effects are a serious problem in spatial statistics because the number of boundary points increases with the dimension. To correct the bias, we may use the estimator based on Guyon (1982):

$$\hat{c}_\eta^g(u, v) = \frac{1}{(n_1 - 2 - u)(n_2 - 2 - v)} \sum_{i=2}^{n_1 - 1 - u} \sum_{j=2}^{n_2 - 1 - v} \eta_{i,j}\eta_{i+u,j+v}.$$

However, Guyon's estimator $\hat{c}_\eta^g$ has some other unpleasant properties. First, it is not guaranteed to be positive definite so that the spectral estimates based on $\hat{c}_\eta^g$ may be negative. Second, the variance of $\hat{c}_\eta^g$ is large, particularly at large lags. To avoid these disadvantages and nevertheless get an asymptotically negligible bias, Dahlhous and Künsch

(1987) used data tapers for the estimation of covariance. The idea of data tapers was originally introduced by Tukey (1967) for nonparametric time series analysis.

Let $q(\cdot)$ be an increasing function in $[0,1]$, twice differentiable, satisfying $q(0) = 0$ and $q(1) = 1$. Define a one-dimensional taper $t(u)$ with the parameter u_N as

$$t(u) = \begin{cases} q(2u/u_N) & (0 \le u < \frac{1}{2}u_N), \\ 1 & (\frac{1}{2}u_N \le u < \frac{1}{2}), \\ t(1-u) & (\frac{1}{2} < u \le 1). \end{cases} \tag{4.27}$$

Here u_N depends on the sample size N. When large sample properties of the estimator of $\boldsymbol{\theta}$ proposed below are considered, we would like to let $u_N \to 0$ and $n_1 u_N \to \infty$ when $n_1 \to \infty$ and $n_2/n_1 \to r$ (a constant). That is, u_N goes to 0 at a rate slower than $O(1/\sqrt{N})$. See Dahlhous and Künsch (1987) for more details.

A common taper in time series analysis is the Tukey-Hanning taper with

$$q(u) = \frac{1}{2}\left\{1 - \cos(\pi u)\right\}.$$

The taper used by Hart (1989) has

$$q(u) = 10u^3 - 15u^4 + 6u^5.$$

Following Dahlhous and Künsch (1987), the tapered covariance estimator is defined as

$$\begin{aligned} \tilde{c}_\eta(u,v) &= \frac{1}{T_{n_1}T_{n_2}} \sum_{i=2}^{n_1-1-u} \sum_{j=2}^{n_2-1-v} \eta_{i,j}\eta_{i+u,j+v} t\left(\frac{i-0.5}{n_1}\right) t\left(\frac{j-0.5}{n_2}\right) \\ &\quad \cdot t\left(\frac{i+u-0.5}{n_1}\right) t\left(\frac{j+v-0.5}{n_2}\right), \end{aligned} \tag{4.28}$$

where

$$T_{n_i} = \sum_{k=2}^{n_i-1} t^2\left(\frac{k-0.5}{n_i}\right), \qquad (i = 1,2). \tag{4.29}$$

It can be shown that the discrete Fourier transformation of $\{\tilde{c}_\eta(s,t)\}$ at (ω_1, ω_2) is

$$\begin{aligned} \tilde{I}_\eta(\omega_1,\omega_2) &= \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} \tilde{c}_\eta(u,v)\exp(-i(u\omega_1 + v\omega_2)) \\ &= \frac{1}{T_{n_1}T_{n_2}} \left| \sum_{k=2}^{n_1-1} \sum_{l=2}^{n_2-1} \tilde{\eta}_{k,l} \exp(-i(k\omega_1 + l\omega_2)) \right|^2, \end{aligned} \tag{4.30}$$

where

$$\tilde{\eta}_{k,l} = \eta_{k,l} t\left(\frac{k-0.5}{n_1}\right) t\left(\frac{l-0.5}{n_2}\right). \tag{4.31}$$

Because of the symmetry of the periodogram and the power spectral density, it is sufficient to consider ω_1, ω_2 that are in $[0, \pi]$. Let $\omega_1^j = 2\pi j/(n_1 - 2)$, $\omega_2^l = 2\pi l/(n_2 - 2)$, $(j = 1, \ldots, [\frac{n_1-2}{2}]$; $l = 1, \ldots, [\frac{n_2-2}{2}])$. To estimate the spatial dependence parameter $\boldsymbol{\theta}$, we can then choose $\boldsymbol{\theta}$ to maximize the Whittle likelihood

$$L_{n_1,n_2}(\boldsymbol{\theta}) = -\sum_{j=\tau_1}^{[\frac{n_1-2}{2}]} \sum_{l=\tau_2}^{[\frac{n_2-2}{2}]} \left\{\log S_\eta(\omega_1^j, \omega_2^l, \boldsymbol{\theta}) + \tilde{I}_\eta(\omega_1^j, \omega_2^l)/S_\eta(\omega_1^j, \omega_2^l, \boldsymbol{\theta})\right\}. \tag{4.32}$$

where $\tau_1 > 0$ and $\tau_2 > 0$ are tuning parameters.

However, in practice, $\{\eta_{i,j}\}$ are not observable. Hence $\tilde{I}_\eta(\omega_1^j, \omega_2^l)$ can not be calculated. Fortunately, $\{\Delta_Y(i,j)\}$ can be calculated from data, and it is approximately equal to $\{\eta_{i,j}\}$ when the grid is dense enough. Similar to $\tilde{I}_\eta(\omega_1, \omega_2)$, we define

$$\tilde{I}_{\Delta_Y}(\omega_1, \omega_2) = \frac{1}{T_{n_1} T_{n_2}} \left| \sum_{k=2}^{n_1-1} \sum_{l=2}^{n_2-1} \tilde{\Delta}_Y(k,l) \exp(-i(k\omega_1 + l\omega_2)) \right|^2, \tag{4.33}$$

where

$$\tilde{\Delta}_Y(k,l) = \Delta_Y(k,l) t\left(\frac{k-0.5}{n_1}\right) t\left(\frac{l-0.5}{n_2}\right). \tag{4.34}$$

Hence $\tilde{\Delta}_Y(k,l)$ is the tapered difference of the data, and $\tilde{I}_{\Delta_Y}(\omega_1, \omega_2)$ is the periodogram of the tapered difference (of second order). $\tilde{I}_{\Delta_Y}(\omega_1, \omega_2)$ is an approximation of $\tilde{I}_\eta(\omega_1, \omega_2)$. Analogous to the univariate case discussed by Hart (1991), we estimate the spatial dependence parameter $\boldsymbol{\theta}$ by maximizing the approximate Whittle likelihood:

$$\tilde{L}_{n_1,n_2}(\boldsymbol{\theta}) = -\sum_{j=\tau_1}^{[\frac{n_1-2}{2}]} \sum_{l=\tau_2}^{[\frac{n_2-2}{2}]} \left\{\log S_\eta(\omega_1^j, \omega_2^l) + \tilde{I}_{\Delta_Y}(\omega_1^j, \omega_2^l)/S_\eta(\omega_1^j, \omega_2^l)\right\}. \tag{4.35}$$

This is an approximate version of Whittle likelihood (4.32), because $\tilde{I}_{\Delta_Y}(\omega_1, \omega_2)$ is used to approximate $\tilde{I}_\eta(\omega_1, \omega_2)$. The choice of the 2 tuning parameters $\tau_1 > 0$ and $\tau_2 > 0$ affects the quality of the estimate of $\boldsymbol{\theta}$. We must realize that differencing the data does not completely eliminate the effect of the spatial trend (the mean function). If τ_1

and τ_2 are taken too small, the estimate of $\boldsymbol{\theta}$ may be severely biased because of the low frequency contribution of the spatial trend to the periodogram. But τ_1 and τ_2 can not be too large because otherwise we would lose too much information that the periodogram carries at low frequencies.

Let us summarize our method for estimating the covariance of the errors. The key is to find the maximizer of the approximate log-likelihood defined in (4.35). Here we review how to calculate the terms in (4.35). In application, we assume $c_\varepsilon(u, v, \boldsymbol{\theta})$ is the known covariance function of the errors with unknown parameter $\boldsymbol{\theta}$. Given $(\omega_1^j, \omega_2^l, \boldsymbol{\theta})$, $S_\eta(\omega_1^j, \omega_2^l, \boldsymbol{\theta})$ can be evaluated by (4.25) and (4.26). To calculate the tapered periodogram $\tilde{I}_{\Delta_Y}(\omega_1^j, \omega_2^l)$, $\Delta_Y(i, j)$ first needs to be calculated by (4.20). At a given (ω_1^j, ω_2^l), $\tilde{I}_{\Delta_Y}(\omega_1^j, \omega_2^l)$ can then be calculated by (4.33), (4.34), and (4.27). The taper $t(\cdot)$ could be chosen as either the Tukey-Hanning taper or other options.

4.3.2 Simulation study

Again, a simulation study is used to investigate the above method based on differencing and the approximate Whittle likelihood. Consider again the model

$$Y_{i,j} = \sin(2\pi x_{i,1}) + 4(x_{j,2} - 0.5)^2 + \varepsilon_{i,j},$$

where $\varepsilon_{i,j}$ have normal distribution with zero mean, and satisfy

$$\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i+u,j+v}) = \sigma^2 \alpha^{\sqrt{u^2+v^2}} \equiv c_\varepsilon(u, v, \boldsymbol{\theta}),$$

with $\boldsymbol{\theta} = (\sigma, \alpha)^T$ and $0 \le \alpha < 1$. The design points $(x_{i,1}, x_{j,2})$ are on the $n \times n$ grid in the square region (in this case $n_1 = n_2 = n$): $0 \le x_1 \le 1$, $0 \le x_2 \le 1$, with $x_{i,1} = (i - 0.5)/n$ and $x_{j,2} = (j - 0.5)/n$ $(i = 1, \ldots, n; j = 1, \ldots, n)$. In this study, the value of σ is fixed at 0.5. The correlation parameter α can take 3 different values: 0 (no correlation), 0.3 (relatively small correlation), and 0.7 (relatively large correlation). The sample size parameter n can also take 3 different values: 30, 35, and

40. The combination of 3 values of α and 3 values of n gives 9 different cases. For each combination of the values of α and n, 50 independent data sets are simulated. For each data set, 3 values of the tuning parameter τ: 1, 2, and 3 , are investigated (let $\tau_1 = \tau_2 = \tau$). For every τ, we calculate the estimates for the spatial dependence parameters, $\hat{\sigma}$ and $\hat{\alpha}$, by maximizing the approximate log-likelihood (4.35). Note that when calculating $\tilde{I}_{\Delta_Y}(\omega_1, \omega_2)$, the Tukey-Hanning taper with the smoothness parameter $u_N = 0.1$ is used throughout this study (Hart (1991) also used $u_N = 0.1$ in his simulation study for the univariate case, but u_N can be chosen as other values, say 0.05). Linear regression with the Epanechnikov Kernel is used to estimate the spatial trend. Again, two different approaches are used to search bandwidth matrices. The first approach is to consider diagonal bandwidths $\boldsymbol{H} = \text{diag}(h_1, h_2)$, with h_1 and h_2 taking values from 0.1, 0.2, 0.3, 0.4, and 0.5. The second approach is to consider spherical bandwidths $\boldsymbol{H} = h\boldsymbol{I}_2$ only, with h taking values from 0.1, 0.2, 0.3, 0.4, and 0.5. After obtaining $\hat{\sigma}$ and $\hat{\alpha}$ by the method of differencing and approximate Whittle likelihood, the GCV criterion, the corrected C_L criterion, and the corrected GCV criterion can then be evaluated at all candidates of $\boldsymbol{H}$. Hence the $\boldsymbol{H}$ that minimizes each of these criteria can be obtained. The results of this simulation study are summarized in tables 4.10 to 4.14.

Table 4.10 gives the means and the standard deviations of the estimates of σ and the estimates of α from 50 simulations. When α and the tuning parameter τ are fixed, estimates from a a finer grid (larger n) have smaller bias and a smaller standard deviation. When α and the size of the grid n are fixed, the impact of τ is evident. For example, when $\alpha = 0.3$ or 0.7, if τ increases, both the estimates of σ and the estimates of α have a smaller bias but a larger variation. The reason might be, when τ is smaller, more lower frequencies are included in the estimation procedure based on the approximate Whittle likelihood (4.35). Thus the lower frequency components of the spatial trend causes a larger bias. When τ is larger, less frequencies are considered, thus a smaller number of periodogram terms in (4.35) are used. This causes a larger variation for the estimates of the spatial dependence parameters. Based on this table, it is hard to summarize some general guidelines for choosing the value of the tuning parameter τ. Obviously, the value of α has a huge impact on accuracy of the estimates. If we fix the size of the grid n and the tuning parameter τ and let α increase, the bias and the variance of the estimate of σ will increase. So will the variance of the estimate of α. We see that when data are more strongly positively correlated, it is more difficult to get accurate estimates for the spatial dependence parameters.

Table 4.10 The means and the standard deviations (in parentheses) of the estimates of σ and α.

(σ, α)	grid	τ	$\hat{\sigma}$ $(s_{\hat{\sigma}})$	$\hat{\alpha}$ $(s_{\hat{\alpha}})$
(0.5, 0)	30×30	1	0.5178 (0.0111)	0.0657 (0.0325)
	30×30	2	0.5104 (0.0135)	0.0380 (0.0286)
	30×30	3	0.5100 (0.0160)	0.0363 (0.0265)
	35×35	1	0.5157 (0.0121)	0.0621 (0.0285)
	35×35	2	0.5080 (0.0113)	0.0335 (0.0202)
	35×35	3	0.5072 (0.0136)	0.0296 (0.0191)
	40×40	1	0.5128 (0.0096)	0.0523 (0.0184)
	40×40	2	0.5063 (0.0093)	0.0282 (0.0160)
	40×40	3	0.5058 (0.0117)	0.0247 (0.0150)
(0.5, 0.3)	30×30	1	0.5278 (0.0224)	0.3659 (0.0441)
	30×30	2	0.5139 (0.0307)	0.3344 (0.0629)
	30×30	3	0.5142 (0.0574)	0.3270 (0.1019)
	35×35	1	0.5235 (0.0194)	0.3586 (0.0411)
	35×35	2	0.5092 (0.0219)	0.3277 (0.0464)
	35×35	3	0.5084 (0.0399)	0.3224 (0.0747)
	40×40	1	0.5186 (0.0165)	0.3487 (0.0340)
	40×40	2	0.5081 (0.0179)	0.3258 (0.0355)
	40×40	3	0.5061 (0.0272)	0.3197 (0.0528)
(0.5, 0.7)	30×30	1	0.5840 (0.1694)	0.7375 (0.1010)
	30×30	2	0.5410 (0.2196)	0.6867 (0.1483)
	30×30	3	0.5336 (0.2401)	0.6381 (0.1845)
	35×35	1	0.5532 (0.1094)	0.7309 (0.0743)
	35×35	2	0.5368 (0.1807)	0.6824 (0.1252)
	35×35	3	0.5057 (0.2084)	0.6515 (0.1574)
	40×40	1	0.5514 (0.0952)	0.7346 (0.0601)
	40×40	2	0.5670 (0.1789)	0.7173 (0.1089)
	40×40	3	0.5444 (0.2318)	0.6873 (0.1481)

The results of the diagonal bandwidth search are given in Tables 4.11 and 4.12, which display the average values of the pair (h_1, h_2) selected by the 3 criteria, as well as the average values of the optimal choice of (h_1, h_2) based on the 50 simulations. The optimal (h_1, h_2) is chosen from the 25 possible choices by minimizing (3.24). Here the true values of σ and α, and the true mean function are used in finding the optimal (h_1, h_2). Apparently, when there is no correlation ($\alpha = 0$), the bandwidths selected by the 3 criteria are close. But when data are correlated ($\alpha = 0.3$ or 0.9), the GCV criterion tends to choose smaller bandwidth parameters. Especially when the correlation is larger ($\alpha = 0.7$), the GCV criterion tend to choose $h_1 = 0.1$ and $h_2 = 0.1$, corresponding to the smallest bandwidth region. The corrected C_L criterion and the corrected GCV criterion are similar. The bandwidths chosen by these two criteria are much closer to the optimal choice than the GCV criterion.

Table 4.13 shows the mean average squared error for the estimate of the spatial trend from 50 simulations using the bandwidths selected by the 3 criteria, as well as using the optimal bandwidth. When data are correlated, the corrected C_L and the corrected GCV criteria consistently outperform the GCV criterion. And interestingly enough, the corrected GCV criterion apparently has an edge over the corrected C_L criterion. When data are uncorrelated, the 3 criteria perform equally well. From this table, the choice of the tuning parameter τ does not make a considerable difference. When the grid gets finer, the mean average squared error gets smaller. The magnitude of correlation affects the accuracy of the estimate of the spatial trend. When correlation is smaller, the estimate for the spatial trend is more accurate.

Table 4.14 shows the mean average squared error for the estimate of the spatial trend based on 50 simulations using spherical bandwidths. If we make a comparison between this table and Table 4.13, cell by cell, we find that restricting the search to spherical bandwidths generally yields a larger mean average squared error.

From this simulation study, we have numerically shown the advantage of using the criteria proposed in Chapter 3, even if the spatial correlation parameters can only be roughly estimated.

Table 4.11 The average values of h_1 selected by 3 criteria, and the optimal h_1.

(σ, α)	grid	τ	GCV	corrected C_L	corrected GCV	Optimal
(0.5, 0)	30×30	1	0.1580	0.1900	0.1820	0.2000
	30×30	2	0.1700	0.1880	0.1840	0.2000
	30×30	3	0.1540	0.1800	0.1640	0.2000
	35×35	1	0.1560	0.1740	0.1700	0.2000
	35×35	2	0.1420	0.1620	0.1520	0.2000
	35×35	3	0.1660	0.1780	0.1680	0.2000
	40×40	1	0.1580	0.1740	0.1660	0.2000
	40×40	2	0.1620	0.1700	0.1660	0.2000
	40×40	3	0.1620	0.1700	0.1680	0.2000
(0.5, 0.3)	30×30	1	0.1100	0.1980	0.2000	0.2000
	30×30	2	0.1140	0.1880	0.1840	0.2000
	30×30	3	0.1100	0.1740	0.1800	0.2000
	35×35	1	0.1080	0.1920	0.1900	0.2000
	35×35	2	0.1080	0.1840	0.1820	0.2000
	35×35	3	0.1140	0.1740	0.1740	0.2000
	40×40	1	0.1200	0.1940	0.1940	0.2000
	40×40	2	0.1100	0.1800	0.1800	0.2000
	40×40	3	0.1140	0.1820	0.1880	0.2000
(0.5, 0.7)	30×30	1	0.1060	0.2300	0.2120	0.2000
	30×30	2	0.1040	0.2180	0.1940	0.2000
	30×30	3	0.1020	0.1960	0.1720	0.2000
	35×35	1	0.1020	0.2300	0.2140	0.2000
	35×35	2	0.1020	0.2220	0.2080	0.2000
	35×35	3	0.1040	0.2000	0.1840	0.2000
	40×40	1	0.1040	0.2180	0.2140	0.2000
	40×40	2	0.1040	0.2280	0.2160	0.2000
	40×40	3	0.1000	0.2200	0.2020	0.2000

Table 4.12 The average values of h_2 selected by 3 criteria, and the optimal h_2.

(σ, α)	grid	τ	GCV	corrected C_L	corrected GCV	Optimal
(0.5, 0)	30×30	1	0.2940	0.3040	0.3040	0.3000
	30×30	2	0.2800	0.2860	0.2840	0.3000
	30×30	3	0.2860	0.3080	0.2960	0.3000
	35×35	1	0.2900	0.3160	0.3020	0.3000
	35×35	2	0.2940	0.3000	0.2960	0.3000
	35×35	3	0.2920	0.2940	0.2980	0.3000
	40×40	1	0.3040	0.3040	0.3040	0.3000
	40×40	2	0.2900	0.2980	0.2940	0.3000
	40×40	3	0.3000	0.2980	0.3000	0.3000
(0.5, 0.3)	30×30	1	0.1500	0.3620	0.3520	0.3000
	30×30	2	0.1560	0.3480	0.3380	0.3000
	30×30	3	0.1580	0.3480	0.3420	0.3000
	35×35	1	0.1780	0.3620	0.3480	0.3000
	35×35	2	0.1620	0.3340	0.3300	0.3000
	35×35	3	0.1620	0.3280	0.3220	0.3000
	40×40	1	0.2020	0.3520	0.3300	0.3000
	40×40	2	0.1820	0.3140	0.3080	0.3000
	40×40	3	0.1940	0.3140	0.3100	0.3000
(0.5, 0.7)	30×30	1	0.1020	0.4320	0.4240	0.4000
	30×30	2	0.1080	0.3500	0.3620	0.4000
	30×30	3	0.1080	0.3100	0.3300	0.4000
	35×35	1	0.1060	0.4200	0.4040	0.4000
	35×35	2	0.1080	0.3380	0.3500	0.4000
	35×35	3	0.1060	0.3080	0.3140	0.4000
	40×40	1	0.1040	0.4080	0.3920	0.4000
	40×40	2	0.1040	0.3720	0.3720	0.4000
	40×40	3	0.1100	0.3320	0.3360	0.4000

Table 4.13 The mean average squared errors of the estimate of the spatial trend based on 50 simulations using the bandwidths selected by 3 criteria, and using the optimal bandwidths (searching diagonal bandwidths).

(σ, α)	grid	τ	GCV	corrected C_L	corrected GCV	Optimal
(0.5, 0)	30×30	1	0.0222	0.0224	0.0223	0.0220
	30×30	2	0.0233	0.0231	0.0234	0.0221
	30×30	3	0.0230	0.0234	0.0230	0.0225
	35×35	1	0.0206	0.0204	0.0202	0.0197
	35×35	2	0.0221	0.0216	0.0218	0.0205
	35×35	3	0.0214	0.0216	0.0216	0.0206
	40×40	1	0.0214	0.0216	0.0214	0.0200
	40×40	2	0.0214	0.0216	0.0214	0.0200
	40×40	3	0.0225	0.0225	0.0223	0.0212
(0.5, 0.3)	30×30	1	0.0588	0.0329	0.0329	0.0317
	30×30	2	0.0564	0.0345	0.0339	0.0318
	30×30	3	0.0558	0.0364	0.0344	0.0315
	35×35	1	0.0515	0.0330	0.0328	0.0313
	35×35	2	0.0511	0.0322	0.0320	0.0293
	35×35	3	0.0508	0.0335	0.0321	0.0292
	40×40	1	0.0407	0.0285	0.0285	0.0267
	40×40	2	0.0468	0.0275	0.0274	0.0255
	40×40	3	0.0439	0.0285	0.0277	0.0259
(0.5, 0.7)	30×30	1	0.1494	0.1043	0.0997	0.0902
	30×30	2	0.1466	0.1096	0.1025	0.0872
	30×30	3	0.1444	0.1141	0.1067	0.0872
	35×35	1	0.1400	0.0936	0.0902	0.0809
	35×35	2	0.1391	0.1016	0.0942	0.0786
	35×35	3	0.1371	0.1030	0.0983	0.0768
	40×40	1	0.1356	0.0786	0.0773	0.0696
	40×40	2	0.1312	0.0909	0.0832	0.0697
	40×40	3	0.1318	0.1017	0.0909	0.0709

Table 4.14 The mean average squared errors of the estimate of the spatial trend based on 50 simulations using the bandwidths selected by 3 criteria, and using the optimal bandwidths (searching spherical bandwidths only).

(σ, α)	grid	τ	GCV	corrected C_L	corrected GCV	Optimal
(0.5, 0)	30×30	1	0.0231	0.0231	0.0231	0.0231
	30×30	2	0.0230	0.0230	0.0230	0.0230
	30×30	3	0.0232	0.0232	0.0232	0.0232
	35×35	1	0.0213	0.0213	0.0213	0.0213
	35×35	2	0.0217	0.0217	0.0217	0.0217
	35×35	3	0.0216	0.0216	0.0216	0.0216
	40×40	1	0.0210	0.0210	0.0210	0.0210
	40×40	2	0.0220	0.0220	0.0220	0.0220
	40×40	3	0.0222	0.0222	0.0222	0.0222
(0.5, 0.3)	30×30	1	0.0654	0.0391	0.0376	0.0361
	30×30	2	0.0612	0.0377	0.0368	0.0355
	30×30	3	0.0652	0.0413	0.0378	0.0348
	35×35	1	0.0579	0.0361	0.0361	0.0342
	35×35	2	0.0612	0.0336	0.0340	0.0327
	35×35	3	0.0566	0.0354	0.0347	0.0321
	40×40	1	0.0424	0.0297	0.0297	0.0297
	40×40	2	0.0510	0.0295	0.0291	0.0291
	40×40	3	0.0453	0.0310	0.0299	0.0292
(0.5, 0.7)	30×30	1	0.1473	0.1192	0.1174	0.1020
	30×30	2	0.1435	0.1158	0.1097	0.1006
	30×30	3	0.1425	0.1238	0.1177	0.0984
	35×35	1	0.1389	0.1065	0.1016	0.0935
	35×35	2	0.1370	0.1074	0.1064	0.0912
	35×35	3	0.1352	0.1078	0.1036	0.0899
	40×40	1	0.1343	0.0911	0.0868	0.0820
	40×40	2	0.1294	0.0978	0.0920	0.0811
	40×40	3	0.1299	0.1056	0.1001	0.0819

5 CONCLUSIONS

In this dissertation, we discussed kernel type smoothing for spatial data with correlation. The major problem is how to uncover the mean function that describes the relationship between the response variable and a set of predictors from data with correlated errors. Three types of estimators, the Priestley-Chao estimator, the Nadaraya-Watson estimator, and the local linear estimator, have been addressed, with emphasis on the local linear estimator. For correlated data, most former studies on kernel type smoothing had been limited to the case of univariate predictors, with equally spaced design. To our knowledge, we were the first to investigate the more general case: the case of multivariate predictors with random design. We derived formulas for asymptotic mean squared errors of these kernel smoothing estimators, and formulas of asymptotic optimal bandwidths. From these formulas, we can better understand how the correlations affect variances of the kernel smoothing estimators and bandwidth selection. These formulas are useful for further study of "Plug-in" type estimators, which have been extensively discussed for uncorrelated data. In the presence of spatially correlated errors, we have shown that traditional data-driven bandwidth selection methods, such as cross-validation and generalized cross-validation, fail to provide good bandwidth values. We proposed some data-driven bandwidth selection methods that account for the presence of spatial correlation. Simulation studies have shown that these methods are effective when the covariances between errors are completely known. When the covariances need to be estimated from data, we discussed the estimation of the covariances in two special cases: spatial data with repeated measurements, and spatial data collected on a grid (with only one realization). For data with repeated measurements, we proposed an estimation method based on semi-variogram fitting. For data on a grid, we proposed a method based on differencing, and approximate Whittle likelihood estimation. The simulation studies have shown that these methods can provide reasonably good estimates for the purpose of bandwidth selection. Compared to the traditional bandwidth selection criteria which ignore the correlations, our bandwidth selection criteria, based on the estimates of the covariances, can reduce the mean squared error of the estimator for the mean function.

However, it is impossible for us to address all of the related problems in this dissertation. Our work is just a beginning of the application of kernel type smoothing for spatially correlated data. We feel that the following three issues are related and important. First, the asymptotic properties of multivariate kernel smoothing estimators in the case of fixed designs have not been discussed. Second, for spatial data on a grid, the asymptotic properties of our estimators for the spatial dependence parameters based on differencing are still to be investigated. Third, when spatial data are not on a grid, valid methods for estimating the covariances between the errors are still not available. One idea might be transforming the data into grid data by some kind of interpolation. It will be interesting to see how this method works out in the future.

BIBLIOGRAPHY

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, **85**, 749-759.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580-619.

Cheng, M.Y., Fan, J., and Marron, J.S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. *Institute of Statistics Mimeo Series*, #2098, University of North Carolina at Chapel Hill.

Chiu, S. -T. (1989). Bandwidth selection for kernel estimate with correlated noise. *Statistics and Probability Letters*, **8**, 347-354.

Chu, C.-K. and Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, **19**, No. 4, 1906-1918.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association*, **74**, 829-836.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline function: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377-403.

Cressie, N. (1991). *Statistics for Spatial Data*, John Wiley & Sons Inc., New York.

Dahlhaus, R. (1983). Spectral analysis with tapered data. *Journal of Time Series Analysis*, **4**, No. 3, 163-175.

Dahlhaus, R. and Künsch, H. (1987). Edge effects and efficient parameter estimation for stationary random fields, *Biometrika*, **74**, No. 4, 877-882.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*, Dekker, New York.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998-1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics*, **21**, 196-216.

Fan, J. and Gijbels, I. (1992a). *Local Polynomial Modeling and Its Applications*, Chapman Hall, New York.

Fan, J. and Gijbels, I. (1992b). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, **20**, 2008-2036.

Francisco-Fernández, M. and Vilar-Fernández, J. M. (2001). Local polynomial regression estimation with correlated errors. *Working paper*, Departmento de Mathematicas, Facultad de Informatica, Universidad de A Coruna 15071, Spain.

Friedman, J. and Stuetzel, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817-823.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, **19**, 1-141.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, **757** 23-68. Springer-Verlag, New York.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625-633.

Guyon, X. (1982). Parameter estimation for a stationary process on a d-dimensional lattice , *Biometrika*, **69**, No. 1, 95-105.

Hall, P., Lahiri, S.N., and Polzehl, J. (1995). On bandwidth choice in nonparametric regression with both short- and long-range dependent errors. *The Annals of Statistics*, **23**, 1921-1936.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Boston.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, **81**, No. 396, 1080-1088.

Hart, J.D. (1989). Differencing as an approximate de-trending device. *Stochastic Processes and their Applications*, **31**, 251-259.

Hart, J.D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Ser. B*, **53**, No. 1 173-187.

Hart, J.D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society, Ser. B*, **56**, No. 3, 529-542.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, New York.

Herrmann, E., Gasser, T., and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, **79** 783-795.

Herrmann, E. and Wand, M. P. (1995). A bandwidth selector for bivariate kernel regression. *Journal of the Royal Statistical Society, Ser. B*, **57**, No. 1, 171-180.

Masry, E. (1995). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, **17**, No. 6, 571-599.

Masry, E. (1996). Multivariate regression estimation, local polynomial fitting for time series. *Stochastic Processes and Their Applications*, **65**, 81-101.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, **24**, 165-179.

Morris, M. D. and Ebey, S. F. (1984). An interesting property of the sample mean under first-order autoregressive model. *The American Statistician*, **38**, No. 2, 127-129.

Müller, H.-G. (1988), *Nonparametric Analysis of Longitudinal Data*, Springer, Berlin.

Müller, H.-G and Stadmüller (1987), Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, **15**, 610-625.

Müller, H.-G and Stadmüller (1988), Detecting dependencies in smooth regression models. *Biometrika*, **75**, 639-650.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Application*, **9**, 141-142.

Opsomer, J.-D. (1997). Nonparametric regression in the presence of Correlated Errors. *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*, ed. Gregoire T.G., etc., Springer, New York, 339-348.

Opsomer, J.-D., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, **16**, (2), To appear.

Parzen, E. (1959). Statistical inference on time series by Hilbert space methods. I. *Technical Report 23(NR-C42-993), Stanford University, Department of Statistics.*

Parzen, E. (1961). Regression analysis of continuous parameter time series. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory*, ed. Neyman, J., Berkeley: University of California Press. 469-489.

Priestley, M.B. (1994). *Spectral Analysis and Time Series*, Academic Press, San Diego.

Priestley, M.B. and Chao, M.T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society, Ser. B*, **34**, 384-392.

Quintela, A. (1994). A plug-in technique in nonparametric regression with dependence. *Communications in Statistics-Theory and Methods*, **23**, No. 9, 2581-2603.

Rice, J. (1979). On the estimation of the parameters of a power spectrum. *Journal of Multivariate Analysis*, **9**, 378-392.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, **92**, 1049-1062.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22, 1346-1370.

Stein, M. L. (1995), Fixed domain asymptotics for spatial periodograms. *Journal of the American Statistical Association.* **90**, No. 432, 1277-1288.

Stone, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595-645.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric regression. *The Annals of Statistics*, **8**, 1348-1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **13**, 689-705.

Tukey, J. W. (1967). An introduction to the calculations of numerical spectrum analysis. *Advanced Seminar on Spectral Analysis of Time Series*, Ed. Harris, B., 25-46. New York, Wiley.

Wahba, G. (1990). *Spline Models For Observational Data*, SIAM, Philadelphia.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya Ser. A*, **26**, 359-372.

ACKNOWLEDGMENTS

The author wishes to express his deep appreciation to Dr. Jean Opsomer and Dr. Kenneth Koehler for their guidance and helpful discussions. Appreciation is also given to Dr. Soumendra Lahiri, Dr. Yuhong Yang, and Dr. Jay Breidt for their helpful suggestions.