Non-Parametric Estimation of Geographical Relative Risk Functions

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Palmerston North

New Zealand

W. T. P. Sarojinie Fernando

December 2012

Abstract

The geographical relative risk function is a useful tool for investigating the spatial distribution of disease based on case and control data. The most common way of estimating this function is using the ratio of spatial kernel density estimates constructed from the locations of cases and controls respectively. This technique is known as the density ratio method. The performance of kernel density estimators depends on the choice of kernel and the smoothing parameter (bandwidth). The choice of kernel is not critical to the statistical performance of the method but the bandwidth is crucial. Different bandwidth selectors such as least squares cross validation (LSCV) and likelihood cross validation (LCV) are chosen to control the degree of smoothing during the computation of the density ratio estimator.

An alternative way of estimating this relative risk function is local linear regression approach. This deserves consideration since the density ratio estimator can be less natural when the relative risk has a global trend, as one might expect to see when there is a line source of risk such as a polluted river or a road. The use of local linear regression for estimation of log relative risk functions *per se* has not been examined in any detail in the literature, so our work on this methodology is a novel contribution. A detailed account of local linear approach in the estimation of log relative risk function is provided, consisting of an analysis of asymptotic properties and a method for computing tolerance contours to emphasize the regions of significantly high risk. Data driven bandwidth selectors for the local linear method including a novel plug-in methodology is examined.

A simulation study to compare the performance of density ratio and local linear estimators using a range of data-driven bandwidth selectors is presented. The analysis of two specific data sets is examined.

The estimation of the spatial relative risk function is extended to spatio-temporal estimation through the use of suitable temporal kernel functions, since time-scale is an important consideration when estimating disease risk. The extended version of the kernel density estimation is applied here to compute the unknown densities of the spatio-temporal relative risk function. Next we investigate the time derivatives of the space-time relative risk function to see how the disease change with time. This discussion provides novel contributions with the introduction to time derivatives of the relative risk function as well as asymptotic methods for the computation of tolerance contours to highlight subregions of significantly elevated risk. LSCV and subjective bandwidths are used to compute these estimators since it performs well in density ratio method. The analysis on a real application to foot and mouth disease (FMD) of 1967 outbreak is employed to illustrate these estimators.

The relative risk function is investigated when the data include a spatially varying covariate. The discussion produces the introduction to generalized relative risk function in two ways and also asymptotic properties of estimators for both cases as novel works. Generalized kernel density estimation is used to replace the unknown densities in the relative risk function. Asymptotic theories are used to compute tolerance contours to identify the areas which show high risk. LSCV bandwidth selector is described in this estimation process providing the implicit formulae. We illustrate this methodology on data from the 2001 outbreak of FMD in the UK, examining the effect of farm size as a covariate.

Acknowledgements

I am sincerely and heartily grateful to my supervisor, Martin Hazelton, for his continuous support and guidance throughout this research. I am sure this would have not been accomplished without his help. I would also like to thank Martin for being patient specially when I was making less progress. Finally I would like to express my heartfelt gratitude to Martin for his valuable comments and suggestions to direct this dissertation a successful one.

I am thankful to Ganes Ganesalingam for his continuous guidance in numerous ways as being my co-supervisor in the early years of my research. My thanks go to Jonathan Marshall for being my co-supervisor in my final year due to the replacement of Ganes.

I would like to acknowledge all the members in the Statistics group for their valuable comments, and also colleagues in the Institute office of Fundamental Sciences. My special thanks go to IFS for providing me a scholarship at the end of Graduate assistant position and also to Steve Haslett for providing me funds to pay tuition fee afterwards. I should thank my postgraduate colleagues at the Statistics group, Massey University for their encouragement and providing me a pleasing environment.

My special thanks go to two important people. Sarath Kulatunga, a Senior Professor at the University of Kelaniya in Sri Lanka, helped me to acquire a thorough statistical knowledge while I was working towards my M.Phil. with him. I would also like to express my gratitude to Priyantha Wijayatunga, a lecturer at Umea University

in Sweden, who encouraged me to apply to this Graduate assistant Ph.D. position. I take this opportunity to thank to all staff members at the Department of Mathematics, University of Kelaniya. My thanks also go to staff members at the faculty of Applied sciences, specially to my colleagues at the Department of Mathematics, Wayamba University of Sri Lanka for making a friendly environment while I was working there as a lecturer.

Of course, I am indebted to my lovely husband Kithsiri Fernando for his love, support, motivation and constant patience which have taught me so much about sacrifice during this journey. Without him this work would never have come into existence. I am also grateful to my son, Kaveesha Fernando and my daughter, Gihara Fernando for their profound understanding during this period. I missed them so much. I am extremely sorry for the time we spent apart.

I would like to acknowledge my lovely mum, Janet for loving, encouraging me always, believing in me, in all my efforts. I also express my gratitude to my late dad, Julian for making a dream in my mind to be a successful academic when I was a child. I also like to acknowledge my brother and sister for their continuous support. My special remind goes to my mother-in-law, Mary who unfortunately passed away while I was reading for my Ph.D.

Last but not least, I would like to thank my ever loving God for helping and guiding me to be successful throughout my life.

Table of Contents

List of Figures

List of Tables

Notation

$\mathbf{I}$	2×2 identity matrix.
f	Case density function
g	Control density function
r	Relative risk function
ρ	Log relative risk function
$supp(f)$	Support of f
p	Degree of the polynomial, conditional probability and size of covariate.
h	Fixed bandwidth
h_1	Fixed bandwidth for case data.
h_2	Fixed bandwidth for control data.
h_{OS}	Over smoothing bandwidth.
q	Edge correction factor.
q_1	Edge correction factor based on case data.
q_2	Edge correction factor based on control data.
$\mathbf{H}$	bandwidth matrix
Σ	Covariance matrix.
λ	temporal bandwidth
K	Unscaled kernel function
K_h	Scaled kernel function, scaled with h.
$\mu_2(K)$	Second central moment of K
$R(f)$	$\int_{\mathcal{R}} f(\mathbf{x})^2 d\mathbf{x}$
$R(K)$	$\int_{\mathcal{R}} K(\mathbf{z})^2 d\mathbf{z}$
$R(f'')$	$\int_{\mathcal{R}} f''(\mathbf{z})^2 d\mathbf{z}$
f''	Second derivative of the density f.
L	Univariate probability density function.
$\mathcal{D}_{\{}(\mathbf{x})$	First derivative of f with respect to $\mathbf{x}$.
$\mathcal{H}_{\{}(\mathbf{x})$	Hessian of f with respect to $\mathbf{x}$.
$\phi_\sigma(\mathbf{x} - \mu)$	Multivariate normal density.
μ	Mean.

σ^2	Variance.
$\mathcal{R}$	Geographical region
$\|\mathcal{R}\|$	Area of the geographical region $\mathcal{R}$.
∇^2	Laplacian operator.
DR	Density Ratio
LL	Local Linear
RR	Relative risk
ISE	Integrated Squared Error
MSE	Mean Squared Error
AMSE	Asymptotic Mean Squared Error
MISE	Mean Integrated Squared Error
WMISE	Weighted Mean Integrated Squared Error
MIAE	Mean Integrated Absolute Error
AMISE	Asymptotic Mean Integrated Squared Error
PI	Plug-In bandwidth selector
GAM	Generalized additive model.
LSCV	Least Squares Cross Validation
LCV	Likelihood Cross Validation
SE	Standard Error
x_i	scalar.
$\mathbf{x}_i$	A vector of dimension 2
$x_1, ..., x_n$	Random scalar sample of size n.
$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$	a random sample of size n
t	Time of occurence
T	Time interval
$\|T\|$	Length of time interval
FMD	Foot and mouth disease
KDE	Kernel density estimate
$\hat{f}(x;h)$	Fixed kernel density estimate constructed from univariate data x.
$\hat{f}(\mathbf{x};h)$	Fixed kernel density estimate constructed from bivariate case data $\mathbf{x}$.
$\hat{g}(\mathbf{x};h)$	Fixed kernel density estimate constructed from bivariate control data $\mathbf{x}$.
$\hat{r}(\mathbf{x};h)$	Fixed bandwidth relative risk estimate.
$\hat{\rho}(\mathbf{x};h)$	Fixed bandwidth log relative risk estimate.
$\hat{\rho}_{LL}$	LL estimator of log relative risk function.
$\hat{\rho}_{LC}$	Local constant estimator.
$\hat{f}^{-i}(\mathbf{x};h)$	Leave-one-out estimate based on case data
$\hat{g}^{-i}(\mathbf{x};h)$	Leave-one-out estimate based on control data
h_{opt}	The optimal bandwidth
h_{MISE}	MISE- optimal bandwidth
$\hat{h}_{PI}$	Plug-in bandwidth selector
$\hat{h}_{LSCV}$	LSCV bandwidth selector
$\hat{h}_{LCV}$	LCV bandwidth selector
$\hat{h}_{LL.LCV}$	LCV bandwidth for LL estimator.
$\hat{h}_{LL.PI}$	PI bandwidth for LL estimator.

$\hat{h}_{DR.CV}$	LSCV bandwidth for DR estimator.
$\hat{g}_p$	Pooled estimator.
n_1	Case sample size
n_2	Control sample size
n	Equal to $n_1 + n_2$
π	Equal to $\frac{n_1}{n_1+n_2}$
$\bar{L}$	The likelihood function
$\bar{L}'$	The first derivative with respect to β
$\bar{L}''$	The second derivative with respect to β
P, Q	Polynomials.
$p(\mathbf{x})$	Conditional probability for a given point $\mathbf{x}$.
y_i	Binary variable.
Cov	Covariance.
tr	Trace.
δ	A small positive number.
δ_0	A small positive number.
ϵ	A small constant.
$\bar{f}$	Edge corrected kernel density estimate constructed from case data.
$\bar{g}$	Edge corrected kernel density estimate constructed from control data.
H_0,H_0^1,H_0^2	Null hypotheses.
H_1,H_1^1,H_1^2	Alternative hypotheses.
$\mathbb{E}$	Expectation.
θ	A parameter.
Z	Test statistic.
$W_{\mathbf{x}}$	$n \times n$ diagonal matrix at $\mathbf{x}$.

Chapter 1

Introduction

1.1 Motivation and problem description

Modern epidemiology began with the spatial analysis of John Snow's work on cholera in the middle of nineteenth century (see Wand, 2008). However, over the last two decades, there have been major developments in the statistical methods available to investigate the patterns of epidemiological diseases not only in space but also in space and time. So the work in this thesis is motivated by such epidemiological outbreaks in the world, be they in human or animal populations, since it significantly influences to the economy, e.g. foot-and-mouth disease, Salmonella , dengue, cholera, etc. This thesis is concerned with the statistical methods in the field of epidemiology, which can be defined as the study of how disease is distributed among populations and of the factors that impact this distribution. Epidemiology involves the investigation of determinants of human as well as animal health and disease, for the improvement of health care and the prevention of illness. In recent years, epidemiology has become an increasingly important approach in both public health and clinical practice. It is now used together with laboratory research to identify risk factors for diseases.

More specifically, this thesis focuses on statistical methodologies in geographical epidemiology, which is defined as the study of geographical factors affecting the health and illness of populations. The typical characteristic is that the geographical location is considered as an important explanatory variable, either because it reflects an environmentally determined element of risk or because people with similar risk characteristics live together. So this risk varies from place to place. (See Gordis, 2000.)

One of the milestones in the development of epidemiology was the case-control study, which has played a major role during the second half of the twentieth century in identifying the association between the disease and the potential risk factors. It works by taking separate samples of diseased cases and of controls at risk of developing disease (see Gordis, 2000). In such a study, it is important to know about selecting the sample of cases and controls.

Cases can be chosen from different kinds of sources, which include hospital patients, patients in physicians' practices, or clinic patients. Many health authorities maintain registries of patients with certain diseases, such as cancer, which can serve as valuable sources of cases for such studies.

Choice of the most suitable control group is one of the most difficult debated aspects of the study region. One can interpret selecting the control group in different ways. However, the fundamental issue relating to the selection of controls is that they ought to be a representative of all persons without the disease in the population from which

the cases are selected. As an example, controls may be selected either from non-hospitalized people living in the society or from hospitalized patients admitted for diseases other than that for which the cases are admitted.

Generally, case-control studies are best suited to the study of rare diseases (e.g. larynx cancer) and especially useful when a study must be done quickly and inexpensively. Suppose that we have case-control data for a particular disease and we need to analyze them to serve as a measure of how the risk varies from location to location. This amounts to comparing the density function for the incident cases with that of the control population. The natural way to make this comparison is through the ratio, which defines a relative risk function. Bithel (1990, 1992) proposed that relative risk function over a geographical region can be defined as a ratio of two densities, where the numerator represents the case data while the denominator determines the control data.

Let us familiarize with a case-control study. The data used in the Figure 1.1 consists of the residential locations of all the cases of larynx and lung cancer, diagnosed in part of Lancashire, England, during 1974-1983. The locations of 58 larynx cancer are considered as cases while 978 lung cancer as controls since it is standardized by the distribution of controls. The cases and controls are denoted by • and + respectively. An industrial incinerator, which is now disused but is a suspected source of risk for larynx cancer, is also located in this region, denoted by a ■. A lot of cases and controls can be seen in highly populated urban areas while there is low population density in rural areas. The main aim here is to examine the variation of risk over the

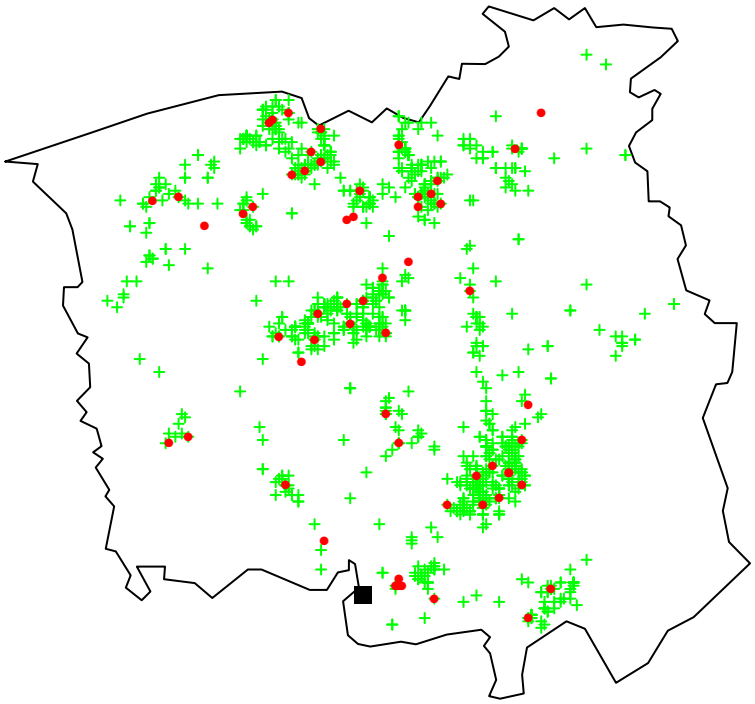

Figure 1.1: The geographical distribution of larynx (58 cases-•) and lung (978 controls-+) cancer data. Incinerator is displayed as ■.

whole region (in essence and exploratory data analysis). In addition to this aim, we wish to assess any increase of the risk of larynx cancer in the vicinity of the incinerator.

Estimates of the relative risk function have been applied effectively in many situations. For example, Kelsall & Diggle (1995a) discuss certain methodological advances in the estimation of relative risk function for the one-dimensional case and extend this work to the bivariate case as in Kelsall & Diggle (1995b). In the latter paper, they have

introduced tolerance contours in order to assess the significant elevated risk surface. Apart from this methodological research, some have worked on real applications of the relative risk function. That is, Wheeler (2007) carries out a comparison of spatial clustering and cluster detection techniques for childcare leukemia incidence in Ohio. Sabel et al. (2000) study the estimation of relative risk for motor neurone disease in Finland. Kelsall & Diggle (1998) study two methods in the relative risk estimation. They are the density ratio (our traditional method) and the binary regression approach. They analyze cancer data from the Walsall district health authority, UK.

As noted previously, the relative risk function is the ratio of case to control densities. These must be estimated in practice. There are two approaches to determine the unknown density function for a random sample. One can either choose the parametric approach, which requires the assumption that the random sample belongs to a parametric family of distributions and then estimating the unknown parameters, or the non-parametric approach, which requires less rigid assumptions about the density of the random sample (Silverman, 1986). An obvious drawback of the parametric approach is that important data structure can be masked when there is no previous knowledge of the sample to assist in the choice of the parametric family of distributions. So in this thesis, we focus on non-parametric models in our estimators.

1.2 Organization of the thesis

There are three main objectives of this thesis. They are to examine the spatial variation in risk of disease over a geographical region, to identify hot-spot areas of increased

risk and to investigate how this risk changes through time or with other covariates. Note that we focus on methods for preliminary exploration and visualization of the relative risk over space, rather than formal model building and inference.

Chapter 2 commences with a comprehensive review of spatial relative risk function and its estimation using kernel smoothing, which is used to estimate the unknown densities in the relative risk function. These methods are discussed in both univariate and multivariate (bivariate) cases. Moreover, we investigate the crucial matter of selecting the smoothing parameter for kernel density estimation. We then look at edge effects which arise when there are observations near the boundary of the region, and stabilization of the relative risk estimates in areas of sparse data. This Chapter also provides a discussion about the existing asymptotic properties of bias and variance of kernel density estimates and relative risk estimates (on log scale). The use of kernel density estimates (KDE) and relative risk estimates is illustrated in two real applications, to the Chorley-Ribble larynx cancer data that we introduced earlier, and data on Myrtle-Wilt in Tasmania.

Chapter 3 examines two cross-validation bandwidth selectors, namely least squares cross validation (LSCV) and likelihood cross validation (LCV). Then we carry out a simulation study to contrast the performance of LSCV over LCV in the estimation of log relative risk function. The performance of these bandwidth selectors are then illustrated in the Chorley-Ribble larynx cancer data.

We explore a local linear method for estimation of the log relative risk function in

Chapter 4. We provide a direct proof of results for the asymptotic bias and variance of the local linear estimator. A large simulation study is carried out to contrast the performance of the density ratio method over local linear approach in the computation of relative risk function. Here, we introduce a new plug-in bandwidth selection method for the local linear estimator. During the simulation study, LSCV bandwidth selector is used to compute density ratio estimator while LCV and our novel plug-in (PI) bandwidth selectors are employed for local linear estimator. The use of the relative risk estimates is illustrated in three real data sets, Myrtle Wilt, Chorley-Ribble and FMD in 1967 outbreak. These real data are used to compute log relative risk estimates for both methods and highlight the regions of significantly elevated risk by computing 95% tolerance contours. The work in this Chapter 4 is a novel contribution.

In the first half of Chapter 5, we expand the work on spatial analysis in Chapter 2 into the context of space-time. We define a spatio-temporal relative risk function and also a kernel density function. As in spatial analysis, we discuss about LSCV and LCV bandwidth selectors in spatio-temporal relative risk estimation. The formula for the asymptotic properties of spatio-temporal estimators are obtained and then tolerance contours are examined to identify the regions having significantly high risk. In the second half of Chapter 5, we introduce time derivatives of log relative risk function since the rate of change of disease risk is important. The theoretical properties of its estimators are derived. These estimates are illustrated with a real application to foot and mouth disease (FMD) data highlighting the use of 5% tolerance contours to identify the areas where the relative risk is changing. The work in the second half of Chapter 5 retains as a novel contribution to the field of spatio-temporal exploratory

data analysis.

We generalize the relative risk function in Chapter 6 when there is a spatially varying covariate. The relative risk function is defined in two ways. We provide the asymptotic properties in both cases. LSCV is employed to choose the bandwidth. A real application to FMD, which is diagnosed in the county of Cumbria, Northern England in 2001, is used to produce these estimators when the covariate is considered as the total farm population. The whole work we did in Chapter 6 is a novel addition to the relevant fields.

In Chapter 7, we summarize the results, acquire throughout this thesis and suggest future directions that can be carried on.

For convenience, proofs of results are collected together in the Appendix.

Chapter 2

Non-parametric estimation of spatial relative risk function

2.1 Introduction

Suppose that we observe the locations of randomly sampled cases and controls for a particular disease within a designated geographical region $\mathcal{R}$. A question of considerable epidemiological interest is whether or not the disease risk varies across the region, and if so in what manner. To answer this, suppose that we obtain random samples of the locations of n_1 cases and n_2 controls, from densities f and g respectively. We order the data so that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ are case locations and $\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_1+n_2}$ are control locations. Bithell (1990) defined the geographical relative risk function, r for $\mathbf{x} \in \mathcal{R}$ as a density ratio

$$r(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})}. \tag{2.1.1}$$

As pointed out by Wakefield and Elliott (1999), it would be more precise to refer to r as a relative odds function, but the distinction is essentially ignorable if the disease is rare. The mean value of $r(\mathbf{x})$ averaged, with respect to the control density over $\mathcal{R}$ is one i.e.,

$$\begin{aligned} E_{\mathbf{x}\sim g}[r(\mathbf{x})] &= \int_{\mathcal{R}} r(\mathbf{x})g(\mathbf{x})\ d\mathbf{x} \\ &= \int_{\mathcal{R}} \frac{f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})\ d\mathbf{x} \\ &= \int_{\mathcal{R}} f(\mathbf{x})\ d\mathbf{x} \\ &= 1. \end{aligned}$$

This means that $r(\mathbf{x}) > 1$ indicates elevated risk of disease at the point $\mathbf{x}$ in comparison to the region as a whole.

In practice, this relative risk function can be estimated by the ratio of kernel density estimates. i.e.

$$\hat{r}(\mathbf{x}) = \frac{\hat{f}(\mathbf{x})}{\hat{g}(\mathbf{x})}\ ;\ \mathbf{x} \in \mathcal{R}. \tag{2.1.2}$$

where $\hat{f}$ and $\hat{g}$ represent the kernel density estimates constructed from case and control data respectively. See for example Bithell (1991). Kelsall & Diggle (1995a) discussed the estimation of r for the one dimensional case and extended to the bivariate setting in Kelsall & Diggle (1995b). In this estimation process, Kelsall & Diggle (1995a,b) suggest using the relative risk estimator on log scale in order to handle f and g symmetrically.

We therefore explicitly define the log relative risk function for $\mathbf{x} \in \mathcal{R}$ as

$$\begin{aligned} \rho(\mathbf{x}) &= \log\left[r(\mathbf{x})\right] \\ &= \log\left[\frac{f(\mathbf{x})}{g(\mathbf{x})}\right] \\ &= \log\left[f(\mathbf{x})\right] - \log\left[g(\mathbf{x})\right] \end{aligned} \tag{2.1.3}$$

and an estimate of it is given by

$$\hat{\rho}(\mathbf{x}) = \log[\hat{f}(\mathbf{x})] - \log[\hat{g}(\mathbf{x})]. \tag{2.1.4}$$

This density ratio methodology has been used successfully in a variety of applications in both human and veterinary epidemiology, see Berke (2005) and Wheeler (2007). In particular, this direct density ratio estimator of the relative risk function has proven adept at identifying approximately circular or elliptical areas of elevated risk, often around postulated point sources of risk (Hazelton & Davies, 2009). Primarily we shall focus on the estimation of ρ rather than r.

The performance of density ratio depends critically on the choice of bandwidth. Progress in the development of reliable data-driven bandwidth selectors for estimation of relative risk has been relatively slow. Kelsall & Diggle (1995a, 1998) examine the least squares and likelihood cross-validation techniques, while Hazelton (2008) has proposed a weighted cross-validation method. Kelsall & Diggle (1998) have conducted a small simulation study including the density ratio estimator implemented using cross-validation bandwidth selectors, and have found relatively little difference in performance. The choice of bandwidth selectors are described in detail in Chapter 3.

In the remainder of this chapter, we review methods and techniques in kernel density estimation that will provide the necessary foundation for the developments in subsequent chapters. We define the kernel density estimator for univariate case in the estimation of spatial relative risk function (Silverman 1986) and provide a brief description of it in Section 2.2.1. Then we extend it to bivariate case in Section 2.2.2. According to previous research, sparse data can be a problem for relative risk estimation. So it is discussed in Section 2.3.1. It is common to observe data near the boundary of the region and this leads to edge effects during density estimation. Edge correction process is examined in Section 2.3.2. We present the fundamental theories for the density ratio approach in Section 2.4. Also we look at the bias and variance asymptotic properties for the kernel density estimator and hence log relative risk estimator. In Section 2.5, we examine the computation of 5% tolerance contours to highlight the areas of significantly elevated risk. We analyse two real applications in Section 2.6 to illustrate the use of these estimators. The first data set relates to the incidence of larynx cancer in the Chorley-Ribble Health region of Lancashire in England while the second is concerned with the distribution of Myrtle Wilt in an area of forest in the Australian state of Tasmania.

2.2 Kernel density estimation

Kernel smoothing is an important data smoothing technique where inferences about the population are made, based on a finite data sample. The simplest data smoothing application is for estimation of probability density functions. Kernel density estimation is a method to estimate the probability density function of a random variable

when a parametric model is inappropriate or unknown, i.e. a non-parametric approach. Most non-parametric density estimators are more computationally demanding and a lot of research on this can be found in the literature.

The earliest non-parametric density estimator for univariate case is the histogram. After that kernel, orthogonal series and nearest neighbor methods were invented. Later, penalized likelihood, polynomial spline, variable kernel were introduced. Histograms, frequency polygons, spline estimators, orthogonal series estimators and penalised likelihood estimators are discussed in Silverman (1986), Scott (1992) and Simonoff (1996). However, we focus on kernel density estimators because they are easy to interpret and to implement. See Wand & Jones (1995).

2.2.1 Univariate case

We begin with the univariate kernel density estimation since it is the most straightforward among several types of kernel estimators. The properties can be studied thoroughly based on its simplicity. Suppose we have a random sample $x_1, \ldots, x_{n_1}$, drawn from a common density f. Then the univariate kernel density function can be estimated by $\hat{f}$ as follows

$$\hat{f}(x; h) = n_1^{-1} \sum_{i=1}^{n_1} K_h (x - x_i) \tag{2.2.1}$$

where K is the unscaled kernel function which is typically is a symmetric probability density function with finite variance. K_h is the scaled kernel function and h is the constant bandwidth which is a positive, non-random number. The scaled and unscaled kernels are related by $K_h(x) = h^{-1}K(h^{-1}x)$. To obtain $\hat{f}(x; h)$, a scaled

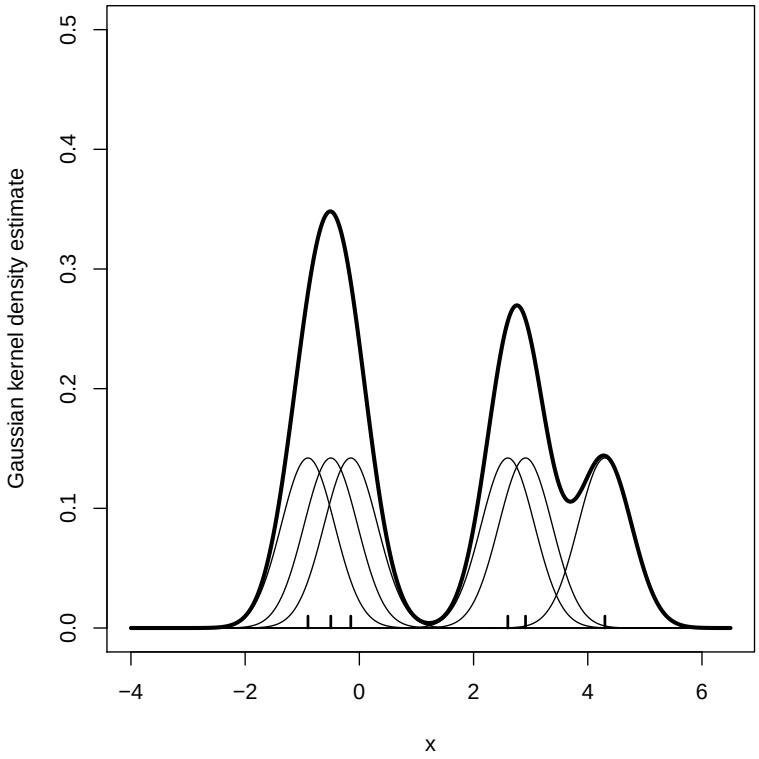

Figure 2.1: Univariate kernel density estimate based on six observations: solid lines - individual kernels, bold line - kernel density estimate

kernel function of probability mass n_1^{-1} is placed at each data point. These are then summed up together to give a combined curve. This combined curve is the kernel density estimate as given in Figure 2.1. Six data points are x_1 = -0.9, x_2 = -0.5, x_3 = -0.15, x_4 = 2.6, x_5 = 2.91 and x_6= 4.3, marked on the x-axis. The kernel K is the standard normal probability density function (the solid lines are the scaled kernels). We see that the kernel density estimate is trimodel, reflecting the shape of the data. Here we use a subjective bandwidth, $h = 0.468$.

The kernel function, K determines the shape of the curve while the bandwidth h

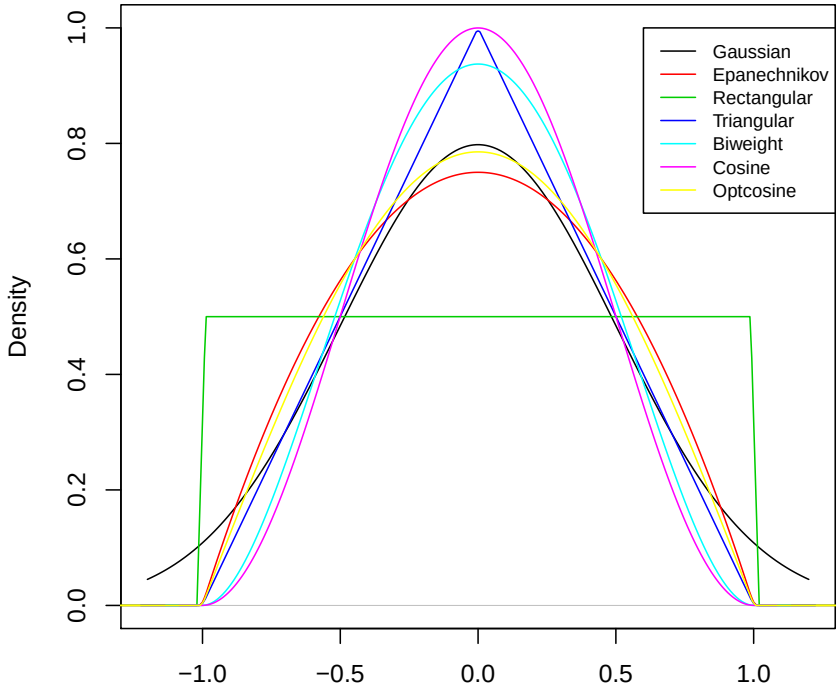

Figure 2.2: Scaled univariate kernel functions.

determines its width. It is generally accepted that the kernel function is not critical (see Silverman 1986, Chap. 3). Some commonly used kernel functions are uniform, triangular, biweight, triweight, Epanechnikov, Gaussian (normal), etc. Figure 2.2 shows the kernel functions that are commonly applied in density estimation. The Gaussian kernel is often used, i.e. $K(x) = \phi(x)$, where ϕ is the standard normal density function.

The choice of bandwidth is particularly important, since it controls the amount of smoothing. The effect of varying bandwidth is given in Figure 2.3 which shows the

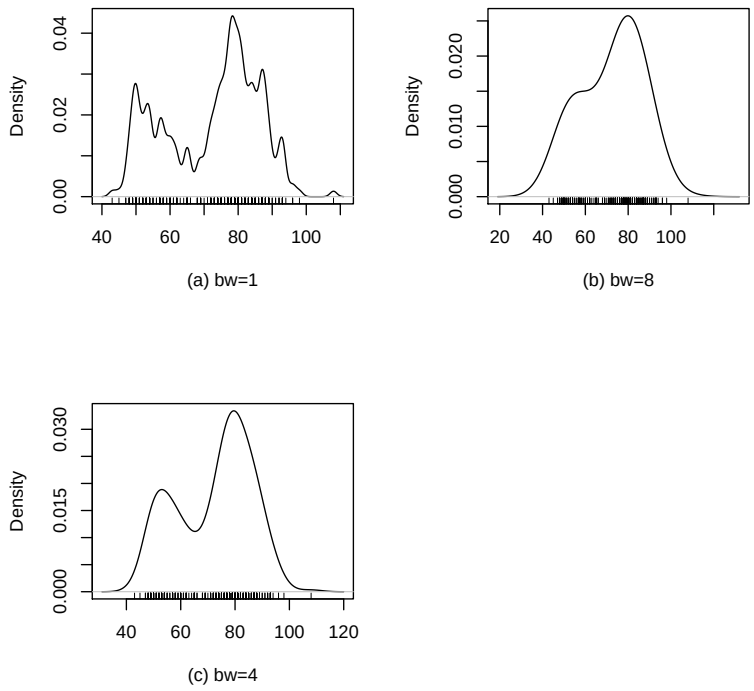

Figure 2.3: Kernel density estimates for 299 observations of a bimodal density. Bandwidths: (a) 1; (b) 8; (c) 4.

kernel density estimates constructed from different bandwidths. The crucial task is thus to find an automatic, data-driven bandwidth selector. The estimates have been constructed from a set of 299 observations which is of eruptions of Old Faithful geyser in Yellowstone National Park, USA (Weisberg, 1980). In Figure 2.3(a), kernel estimate is constructed using bandwidth, $h = 1$. It can be seen that the narrowness of the kernel means that the averaging process performed at each point is based on relatively few observations resulting in a very rough estimate of $\hat{f}(x)$. This estimate pays too much attention to the particular data set at hand and does not allow for the variation across samples. Such an estimate is said to be undersmoothed. Figure

2.3(b) shows the kernel density estimate constructed from the same data set, but with different bandwidth, h=8. This result is a much smoother estimate which is really too smooth since the bimodal structure has been smoothed away. This is an example of an oversmoothed estimate. A compromise is reached with the bandwidth, h=4 in Figure 2.3(c). In this case, the kernel estimate is not overly noisy, yet the essential structure of the underlying density has been recovered which is bimodal. So care must be taken while estimating the smoothing parameter. The choice of bandwidth is discussed in detail in chapter 3. This univariate case is extended to the bivariate setting in the next Section.

2.2.2 Bivariate case

The demand for non-parametric density estimates for recovering the structure in bivariate data is popular since it is easy to understand a perspective view or contour plot of a two-dimensional density function. However, kernel smoothing estimator in higher dimensions requires the specification of many more bandwidth parameters. Scott and Thompson (1983) consider the presentation of four- and five-dimensional densities. For further details, see Silverman (1986) and Wand & Jones (1995). Throughout this thesis, we use bold face to display the locations in higher dimensional space.

Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ is a random sample drawn from a generic density f. Then the single bandwidth 2-dimensional kernel density estimator is defined by

$$\hat{f}(\mathbf{x}; h) = n_1^{-1} \sum_{i=1}^{n_1} K_h \left(\mathbf{x} - \mathbf{x}_i\right). \tag{2.2.2}$$

where $\mathbf{x} = (x_1, x_2)^T$ and $\mathbf{x}_i = (x_{i1}, x_{i2})^T, i = 1, ..., n_1$. Here K is the unscaled kernel,

K_h is the scaled kernel and h is the bandwidth. The scaled and unscaled kernels are linked by $K_h(\mathbf{x}) = h^{-2}K(h^{-1}\mathbf{x})$ and K is a bivariate kernel function satisfying

$$\int_{\mathcal{R}} K(\mathbf{x})d\mathbf{x} = 1 \tag{2.2.3}$$

(see Cacoullos, 1966). A well-known choice for kernel K is the standard bivariate Gaussian density function

$$K(\mathbf{x}) = (2\pi)^{-1}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right) \tag{2.2.4}$$

in which case $K_h(\mathbf{x} - \mathbf{x}_i)$ determines the normal distribution with mean $\mathbf{x}_i$ and covariance matrix h^2I.

We concentrate our attention on kernel functions K that are spherically symmetric probability density functions so that the scaled kernel K_h is isotropic. This is arguably desirable for applications in geographical epidemiology. See Wand & Jones (1993) and Doung & Hazelton (2003). As they lead to smooth density estimates and simplify the mathematical analysis, we use Gaussian kernels throughout this thesis. In this Section, we are considering bivariate kernel density estimators since our data is (primarily) bivariate.

One of the fundamental steps in exploration of a bivariate data set is to examine a scatter plot of the data. However, it is often the case that a density estimate will detect features which cannot be seen in the scatter plot (Silverman, 1986). An example is given in Figure 2.4. These data contains 58 larynx cancer bivariate data from the Chorley-Ribble health authority. The dispersion of this data set is already shown in the Figure 1.1 of Chapter 1. The left panel of Figure 2.4 shows the scatter plot of

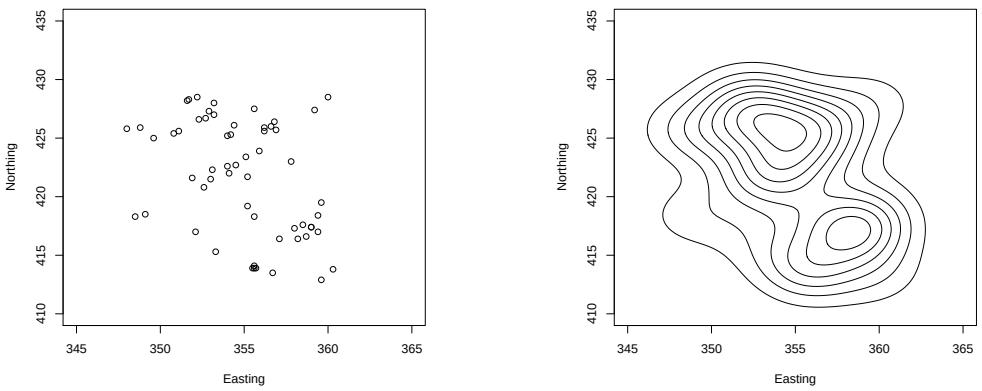

Figure 2.4: Left panel displays the scatter plot of larynx cancer data and the right panel shows the bivariate kernel density estimate, constructed from these data. The subjective bandwidth, 2 is used.

larynx cancer data and the right panel displays the bivariate kernel density estimate constructed from these data. Here in the latter panel, two modes are visible and these modes indicate major regions of having larynx cancer. Note that these modes are not immediately apparent in the scatter plot. Even if the clusters were clearly depicted in the scatter plot, the density estimate would still have the advantage of specifying estimates of the locations of the two modes.

We focus on the fixed bandwidth case in this thesis because there are many unanswered questions about relative risk estimation with fixed bandwidths. Nonetheless, adaptive estimation is also possible (see Davies & Hazelton, 2010).

2.3 Technical problems arising in relative risk estimation

In this Section, we discuss about two problems arising in relative risk estimation. The first problem is regarding sparse data. The relative risk estimator is unstable in areas of sparse data and is undefined if the control density is estimated as zero. We may therefore consider adding a stabilizing constant to the top and bottom of kernel density estimates to avoid such noisy bumps. This is discussed in Section 2.3.1. The boundary effects is the second problem. When the data lie at the boundary of the region, some proportion of KDE is lost and hence make a significant difference to the resultant KDE and hence relative risk estimator. So we use an edge correction technique to control this problem in estimating KDE at boundary points and it is described in Section 2.3.2. These two problems are discussed for bivariate case.

2.3.1 Data scarcity

Sparse data is often encountered in epidemiological studies since the distribution of data can be sometimes highly non-uniform over the geographical region. We recall here the Figure 1.1 of Chorley-Ribble cancer data. There is highly populated urban areas as well as low population density in rural areas. The marked regions in Figure 2.5 shows the sparsely populated rural areas, which are located in southeast and southwest of the geographical region. At these locations, kernel density estimates of cases and controls can be very small or even zero (if using a kernel with finite support), and hence the relative risk estimate is unstable or even undefined. This type of situation requires a bigger bandwidth h in order to control the noise where the

data are sparse. See Hazelton & Davies (2009). Therefore, the authors Bithel (1990) and Anderson & Titterington (1997) suggest a stabilizing process of the underlying density estimators by replacing $\hat{r}$ with

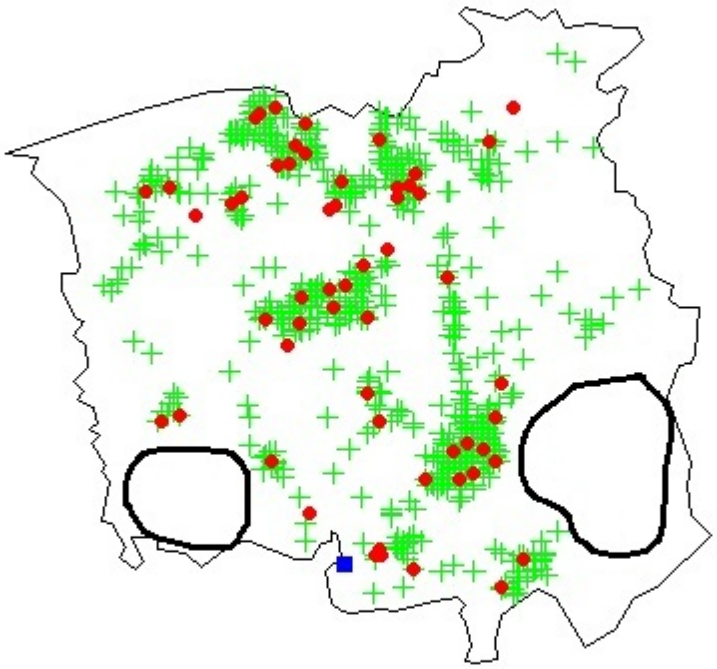

Figure 2.5: Sparse data can be seen mainly in southwest and southeast regions.

$$\tilde{r}(\mathbf{x}) = \frac{\hat{f}(\mathbf{x}) + \delta}{\hat{g}(\mathbf{x}) + \delta} \ , \tag{2.3.1}$$

where $\delta > 0$ is a small positive number. This particular number is labeled as the 'stabilization constant'. Then the log relative risk function can be redefined as

$$\begin{aligned}\tilde{\rho}(\mathbf{x}) &= \log[\tilde{r}(\mathbf{x})] \\ &= \log\left[\frac{\hat{f}(\mathbf{x}) + \delta}{\hat{g}(\mathbf{x}) + \delta}\right]. \end{aligned} \tag{2.3.2}$$

One can choose $\delta = 10^{-8}$ (e.g.). Hazelton & Davies (2009) have chosen $\delta = \delta_0 \max_{\mathbf{x}}[\hat{g}(\mathbf{x})]$ for a small constant δ_0. While this is an interesting idea, it is unclear as to how δ may be chosen optimally, and what the effects are on the properties of the relative risk estimator. We do not pursue this methodology further in this thesis, but we choose a large bandwidth to resolve this issue by controlling the variance of the estimates in areas of low density.

2.3.2 Edge correction

All spatial analyses in epidemiology are usually carried out within a finite geographical region $\mathcal{R}$. So our theory on kernel density estimates is based on the assumption that the densities of the cases and controls outside the study region $\mathcal{R}$ are zero. But in practice, it is usually the case that some observations lie at the boundary of the region of interest. The raw estimated densities are positive outside the region, thereby introducing a bias near the edges of $\mathcal{R}$. This can be clearly seen in Figure 2.6. Here the shaded subregions contain mass from the kernel density estimates and hence result in a loss of probability mass within the finite region. Density estimates, and hence relative risk estimates, constructed from such data tend to be misrepresented by these boundary effects. So we should employ boundary corrections methods to dodge this obstacle. Lawson et al. (1999) have mentioned about elimination of edge effects in disease mapping. Some suggestions for reducing edge effects can be found in Marron

& Ruppert (1994). However, many of the one-dimensional edge correction methods in the current literature are not easily extensible to higher dimensions. For this reason, we shall use simple explicit edge correction terms, introduced by Diggle (1985) and later by Kelsall & Diggle (1995b). They give an adjusted kernel estimator as follows:

$$\bar{f}_h(\mathbf{x}) = \frac{\hat{f}_h(\mathbf{x})}{q_h(\mathbf{x})} \ ,$$

where $q_h(\mathbf{x}) = \int h^{-2} K\{h^{-1}(\mathbf{u} - \mathbf{x})\} d\mathbf{u}$ is the correction factor for the case data.

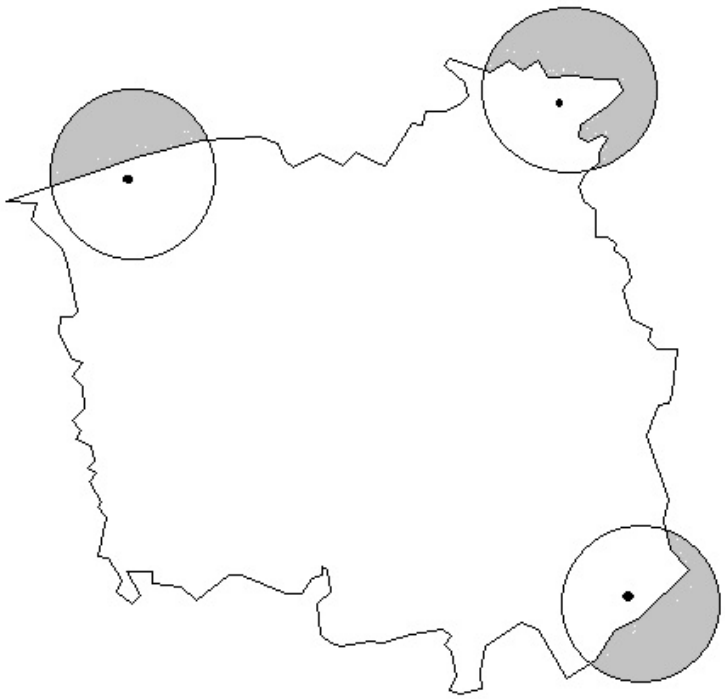

Figure 2.6: Shaded areas represent the subregions which contribute to the kernel estimate, also is located outside of the region, so needs edge correction.

Note that $\bar{f}_h(\mathbf{x})$ does not strictly integrate to unity on $\mathcal{R}$. Similarly, $\bar{g}_h(\mathbf{x})$ can be

defined as above for the control data. Using $\bar{f}$ and $\bar{g}$, we rescale relative risk estimate, $\bar{r}$ including edge correction terms as follows.

$$\begin{aligned}\bar{r}_{h_1,h_2}(\mathbf{x}) &= \frac{\bar{f}_{h_1}(\mathbf{x})}{\bar{g}_{h_2}(\mathbf{x})}; \mathbf{x} \in \mathcal{R} \\ &= \frac{\hat{f}_{h_1}(\mathbf{x})}{q_1(\mathbf{x})} \frac{q_2(\mathbf{x})}{\hat{g}_{h_2}(\mathbf{x})} \\ &= \hat{r}_{h_1,h_2}(\mathbf{x}) \frac{q_2(\mathbf{x})}{q_1(\mathbf{x})} ,\end{aligned}$$

where $q_1(\mathbf{x}) = q_{h_1}(\mathbf{x})$ and $q_2(\mathbf{x}) = q_{h_2}(\mathbf{x})$.

Hazelton (2008) mentions that the rescaling factors cancel out when common bandwidths are chosen for the estimation of case and control densities. That is, $h_1 = h_2$ and then $\bar{r}_{h_1,h_2}(\mathbf{x}) = \hat{r}_{h_1,h_2}(\mathbf{x})$. So we use common bandwidths in this thesis for the case and control density estimates to avoid the need for edge correction.

2.4 Asymptotic properties

2.4.1 Univariate case

Let us consider $\hat{f}(x; h)$ as an estimator of the density function $f(x)$ at some point x and X be a random variable having the density f. Assume that

1. The density f has second derivative f'' which is continuous and square integrable.

2. The bandwidth $h \to 0$ and $nh \to \infty$ as $n \to \infty$ where $n = n_1 + n_2$ with n_1/n_2 fixed.

3. The kernel, K is a bounded probability density function having finite second

moment and symmetric about the origin, i.e. $\int K(x)dx = 1$, $\int xK(x)dx = 0$, $\mu_2(K) = \int x^2 K(x)dx < \infty$.

Assuming that the above conditions are satisfied, it can be shown that (see Bartlett, 1963) the bias and the variance of $\hat{f}(x;h)$ is as follows:

$$\text{Bias}[\hat{f}(x,h)] = \frac{1}{2}h^2\mu_2(K)f''(x) + o(h^2) \tag{2.4.1}$$

and

$$\text{Var}[\hat{f}(x,h)] = n_1^{-1}h^{-1}f(x)R(K) + o(n_1^{-1}h^{-1}), \tag{2.4.2}$$

where $R(K) = \int K(\mathbf{x})^2 d\mathbf{x}$. Notice that the bias of $\hat{f}$ is asymptotically proportional to h^2. So to reduce bias, one needs to take small bandwidth h. However, the bandwidth, h to be small makes a rise in the variance of $\hat{f}$ since the variance of $\hat{f}$ is proportional to $n_1^{-1}h^{-1}$. By adding the variance and the squared bias of $\hat{f}(x,h)$, we get the mean squared error

$$MSE\{\hat{f}(x,h)\} = n_1^{-1}h^{-1}f(x)R(K) + \left[\frac{1}{2}h^2\mu_2(K)f''(x)\right]^2 + o(n_1^{-1}h^{-1}) + o(h^4).$$

If we integrate this over $x \in \mathcal{R}$ then we obtain

$$MISE\{\hat{f}(x,h)\} \approx n_1^{-1}h^{-1}R(K) + \left[\frac{1}{2}h^2\mu_2(K)\right]^2 \int [f''(x)]^2 dx.$$

By differentiating with respect to h and setting the derivative equal to zero, we get the closed form expression of the optimal bandwidth as follows.

$$\begin{aligned} h_{\text{opt}} &= \left[\frac{R(K)}{n_1(\mu_2(K))^2 \int [f''(x)]^2 dx}\right]^{1/5} \\ &= \left[\frac{R(K)}{n_1(\mu_2(K))^2 R(f'')}\right]^{1/5}. \end{aligned} \tag{2.4.3}$$

The expression 2.4.3 is interpretable as that h_{opt} is inversely proportional to $[R(f^{''})]^{1/5}$. Since $|f^{''}(x)|$ is a measure of the curvature of f, the functional $R(f^{''})$ measures the total curvature of f. When $R(f^{''})$ is small, more smoothing is optimal while larger $R(f^{''})$ makes the optimal smoothing smaller. We now use these results of Kelsall & Diggle (1995) to derive the properties of $\hat{\rho}$.

Theorem 2.4.1. *Let x be a fixed element in the interior of support of f. Assume that the conditions (1)-(3) are satisfied. Then*

$$Bias[\hat{\rho}(x,h)] = \frac{1}{2}\mu_2(K)\left[h_1^2\frac{f^{''}(x)}{f(x)} - h_2^2\frac{g^{''}(x)}{g(x)}\right] + o(h_1^2 + h_2^2)$$

and

$$Var[\hat{\rho}(x,h)] = R(K)\left[n_1^{-1}h_1^{-1}f(x)^{-1} + n_2^{-1}h_2^{-1}g(x)^{-1}\right] + o(n_1^{-1}h_1^{-1} + n_2^{-1}h_2^{-1}).$$

(See Kelsall & Diggle, 1995). The proof of this theorem can be found in the Appendix A.1.

Note that the bias approximation of $\hat{\rho}$ is zero if $f = g$ and $h_1 = h_2$. We can assume that $f(\mathbf{x}) > \epsilon$ and $g(\mathbf{x}) > \epsilon$ for all $\mathbf{x} \in \mathcal{R}$, where ϵ is assumed known.

2.4.2 Multivariate case

We obtain the asymptotic properties of bias and variance of $\hat{f}$ and $\hat{\rho}$ for multivariate setting by making the following assumptions.

1. Each entry of the Hessian matrix $\mathcal{H}_f(.)$, which is the square matrix of second-order partial derivatives of f, is piecewise continuous and square integrable.

2. $\mathbf{H} = \mathbf{H}_{n_1}$ is a sequence of bandwidth matrices such that $n_1^{-1}|\mathbf{H}|^{-1/2}$ and all entries of $\mathbf{H}$ approach zero as $n_1 \to \infty$. Also we assume the ratio of the largest and smallest eigenvalues of $\mathbf{H}$ is bounded for all n_1.

3. K is bounded, compactly supported d-variate kernel satisfying $\int K(\mathbf{x})d\mathbf{x} = 1$, $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0$ and $\int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = \mu_2(K)\mathbf{I}$, where $\mu_2(K) = \int x_i^2 K(\mathbf{x})d\mathbf{x}$ is independent of i.

Theorem 2.4.2. *(Multivariate Taylor's Theorem)*
Let f be a d-variate function and α_n be a sequence of $d \times 1$ vectors with all components tending to zero. Also let $\mathcal{D}_f(\boldsymbol{x})$ be the vector of first-order partial derivatives of f and $\mathcal{H}_f(\boldsymbol{x})$ be the Hessian matrix of f, the $d \times d$ matrix having (i,j) entry equal to $\dfrac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})$. Then, assuming that all entries of $\mathcal{H}_f(\boldsymbol{x})$ are continuous in a neighborhood of $\boldsymbol{x}$, we have

$$f(\boldsymbol{x}+\alpha_n) = f(\boldsymbol{x}) + \alpha_n^T \mathcal{D}_f(\boldsymbol{x}) + \tfrac{1}{2}\alpha_n^T \mathcal{H}_f(\boldsymbol{x})\alpha_n + o(\alpha_n^T \alpha_n)$$

(See Wand & Jones, 1995).

Theorem 2.4.3. *If the above conditions are satisfied then*

$$Bias[\hat{f}(\boldsymbol{x}, \boldsymbol{H}) = \tfrac{1}{2}\mu_2(K) tr\{\boldsymbol{H}\mathcal{H}_f(\boldsymbol{x})\} + o\{tr(\boldsymbol{H})\}$$

and

$$Var[\hat{f}(\boldsymbol{x}, \boldsymbol{H})] = n_1^{-1}|\boldsymbol{H}|^{-1/2} R(K) f(\boldsymbol{x}) + o(n_1^{-1}|\boldsymbol{H}|^{-1/2}).$$

Theorem 2.4.4. *If the above conditions are satisfied then*

$$Bias[\hat{\rho}(\boldsymbol{x}, \boldsymbol{H})] = \tfrac{1}{2}|\boldsymbol{H}|^{1/2}\mu_2(K)\left[\tfrac{\nabla^2 f(\boldsymbol{x})}{f(\boldsymbol{x})} - \tfrac{\nabla^2 g(\boldsymbol{x})}{g(\boldsymbol{x})}\right] + o(|H|^{1/2})$$

and

$$Var[\hat{\rho}(\boldsymbol{x}, \boldsymbol{H})] = \tfrac{R(K)}{|\boldsymbol{H}|^{1/2}}[n_1^{-1} f(\boldsymbol{x})^{-1} + n_2^{-1} g(\boldsymbol{x})^{-1}] + o\left(n_1^{-1}|\boldsymbol{H}|^{-1/2} + n_2^{-1}|\boldsymbol{H}|^{-1/2}\right).$$

where ∇ is the Laplacian operator.

Since we are considering isotropic smoothing, we use $\mathbf{H} = h^2\mathbf{I}$. Then the bias and variance terms in Theorem (2.4.3) reduces to the following Theorem.

Theorem 2.4.5. *If $\boldsymbol{x}$ is an interior spatial location of $\mathcal{R}$, then*

$$Bias[\hat{f}(\boldsymbol{x}; h)] = \frac{1}{2}\mu_2(K)h^2\nabla^2 f(\boldsymbol{x}) + o(h^2) \tag{2.4.4}$$

and

$$Var[\hat{f}(\boldsymbol{x}, h)] = n_1^{-1}h^{-2}R(K)f(\boldsymbol{x}) + o(n_1^{-1}h^{-2}).$$

The bias and variance estimates of $\hat{\rho}$ in Theorem (2.4.4) becomes

Theorem 2.4.6. *If $\boldsymbol{x}$ is an interior spatial location of $\mathcal{R}$, then*

$$Bias[\hat{\rho}(\boldsymbol{x}; h)] = \frac{1}{2}h^2\mu_2(K)\left[\nabla^2 f(\boldsymbol{x})/f(\boldsymbol{x}) - \nabla^2 g(\boldsymbol{x})/g(\boldsymbol{x})\right] + o(h^2). \tag{2.4.5}$$

and

$$Var[\hat{\rho}(\boldsymbol{x}, h)] = \frac{R(K)}{h^2}\left[n_1^{-1}f(\boldsymbol{x})^{-1} + n_2^{-1}g(\boldsymbol{x})^{-1}\right] + o(n_1^{-1}h^{-2} + n_2^{-1}h^{-2}). \tag{2.4.6}$$

The proof of the approximations in (2.4.5) and (2.4.6) are given in the Appendix A.2. Note that the bias approximation of $\hat{\rho}$ is zero if $f = g$, i.e., $\hat{\rho}$ is an unbiased estimator of ρ. However, we should avoid the possibility that $\hat{g} = 0$ (or $\hat{f} = 0$), when estimating

ρ. Combining the terms of integrated variance and integrated squared bias of $\hat{f}$ to obtain MISE,

$$\begin{aligned} MISE\{\hat{f}(\mathbf{x})\} &= \int \mathrm{Var}[\hat{f}] + \int [\mathrm{Bias}(\hat{f})]^2 \\ &\approx n_1^{-1} h^{-2} R(K) + \frac{1}{4} h^4 [\mu_2(K)]^2 \int [\nabla^2 f(\mathbf{x})]^2 d\mathbf{x}. \end{aligned}$$

By differentiating the MISE with respect to h and setting the derivative of it equal to zero,

$$\begin{aligned} h_{\mathrm{opt}}^6 &= 2R(K)[\mu_2(K)]^{-2} n_1^{-1} \left[\int [\nabla^2 f(\mathbf{x})]^2 d\mathbf{x} \right]^{-1} \\ &= \frac{2R(K)[R(\nabla^2 f)]^{-1}}{[\mu_2(K)]^2 n_1}. \end{aligned}$$

The conclusion we draw here is comparable to that for the univariate case. The approximately optimal bandwidth converges to zero as n_1 increases, but does so extremely slowly, at the rate $n_1^{-1/6}$. Furthermore, the appropriate value of h_{opt} depends on the unknown density being estimated, $R(\nabla^2 f)$. Similarly, in terms of $\hat{\rho}$,

$$\begin{aligned} MISE\{\hat{\rho}(\mathbf{x})\} &= \int \mathrm{Var}[\hat{\rho}] + \int [\mathrm{Bias}(\hat{\rho})]^2 \\ &= h^{-2} R(K) \int \left[n_1^{-1} f(\mathbf{x})^{-1} + n_2^{-1} g(\mathbf{x})^{-1} \right] d\mathbf{x} \\ &\quad + \frac{1}{4} h^4 [\mu_2(K)]^2 \int \left[\nabla^2 f(\mathbf{x})/f(\mathbf{x}) - \nabla^2 g(\mathbf{x})/g(\mathbf{x}) \right]^2 d\mathbf{x}. \end{aligned}$$

By differentiating MISE with respect to h and setting the derivative of it equal to zero,

$$h_{\mathrm{opt}} = \left[\frac{2R(K) \int [n_1^{-1} f(\mathbf{x})^{-1} + n_2^{-1} g(\mathbf{x})^{-1}] d\mathbf{x}}{\mu_2(K)^2 \int [\nabla^2 f(\mathbf{x})/f(\mathbf{x}) - \nabla^2 g(\mathbf{x})/g(\mathbf{x})]^2 d\mathbf{x}} \right]^{1/6}$$

These results in terms of $\hat{\rho}$ are comparable to that of univariate case. If n_1/n_2 is fixed, then $h_{\mathrm{opt}} = O(n^{-1/6})$ as is standard in bivariate density estimation. If $f \approx g$ over much of $\mathcal{R}$ then the optimal bandwidth will tend to be very large. Due to the helpful bias cancelation when $f = g$, we use $h_1 = h_2 (= h)$ henceforth.

2.5 Tolerance contours of density ratio estimators

Suppose that we obtain a relative risk estimate $\hat{\rho}(\mathbf{x})$ computed from data $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ from the case density f and data $\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n$ $(n = n_1 + n_2)$ from the control density g, both defined on the region $\mathcal{R}$. We wish to highlight the sub-regions of $\mathcal{R}$ where we have observed significantly elevated risk estimates. One way of producing tolerance contours plots is based on a pointwise p-value surface; see Kelsall & Diggle (1995a). The idea is to investigate if the relative risk is greater than one at each point $\mathbf{x} \in \mathcal{R}$ by testing the null hypothesis $H_0 : \rho(\mathbf{x}) = 0$ against the alternative hypothesis $H_1 : \rho(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{R}$.

Kelsall & Diggle (1995a) discussed calculating pointwise tolerance intervals for one-dimensional data and extended it to the bivariate case in Kelsall & Diggle (1995b). We produce a p-value surface $p(\mathbf{x})$ with respect to the above hypotheses. The areas within $p \leqslant 0.05$ (for example) contour are diagnosed as having significantly raised risk. Plotting tolerance contours on a relative risk surface can be used to distinguish significant features of the log relative risk function, from spurious features. Most of the researchers apply this method. Moreover, the construction of tolerance contours should not be regarded as a formal testing procedure, but a method that highlights areas of risk which needs further investigation. We note that this means that multiple testing issues are essentially irrelevant.

There is a choice of methods to obtain the p-value surface $\{p(\mathbf{x}) : \mathbf{x} \in \mathcal{R}\}$. Kelsall and Diggle (1995a) proposed a Monte Carlo method based on random permutations of the case and control labels of each data point. This is computationally exhaustive

and perhaps produces incorrect resultant tolerance contours where high risk surfaces are shown in regions with no data. However, Hazelton & Davies (2009) developed an alternative asymptotic $z-$test statistic, which is computationally less expensive and provides a more reliable method to obtain tolerance contours.

We therefore construct p-values for constructing tolerance contours for $\hat{\rho}$ by using the z-test test statistic. Since the bias is zero under H_0,

$$\tilde{Z}(\mathbf{x}) = \hat{\rho}(\mathbf{x})/SE\{\hat{\rho}(\mathbf{x})\},$$

where SE is the standard error, given by

$$SE\{\hat{\rho}(\mathbf{x})\} = h^{-1}\sqrt{R(K)[n_1^{-1}f(\mathbf{x})^{-1} + n_2^{-1}g(\mathbf{x})^{-1}]} \tag{2.5.1}$$

Assuming that the null hypothesis is true (under H_0), the two sets of data arise from a common density $f = g$. In that case both f and g in (2.5.1) can be replaced by a pooled estimator, $\hat{g}_p$ constructed from both cases and controls. We note that if $\mathbf{x}$ falls at the boundary of the region, $\mathcal{R}$, then it is necessary to apply edge correction (see Section 2.3.2) to the density estimator $\hat{g}_p$ and also replace $R(K)$ in equation (2.5.1) by $\int_{\mathcal{R}}\{K_h(\mathbf{x})\}^2 d\mathbf{x}$. The results of Hall & Marron (1988) justify a normal approximation to the sampling distribution of $\hat{\rho}(\mathbf{x})$.

2.6 Real applications

In this Section, we consider the estimation of spatial relative risk in two particular real applications.

2.6.1 Chorley-Ribble data

The first example, which is concerned with a well known example on cancers of the larynx and lung diagnosed in the Chorley-Ribble region of Lancashire, England, is already discussed in Chapter 1. This data set was initially analysed and presented by Diggle (1990) and after that subsequently been analysed by Diggle and Rowlingson (1994) and Baddeley et al. (2005).

Diggle (1990) and Diggle & Rowlingson (1994) have shown a significant association

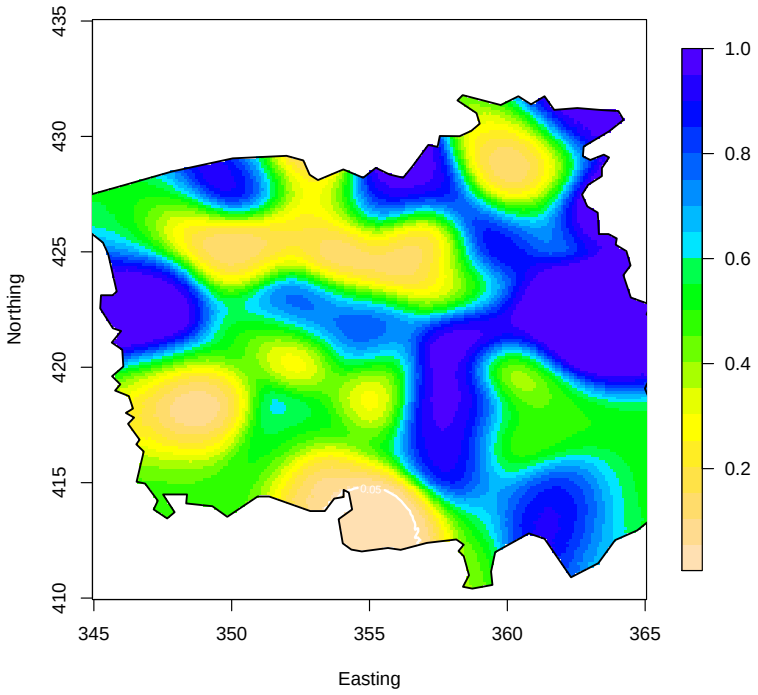

Figure 2.7: P-values surface based on the asymptotic theory describing the excess risk for a given fixed bandwidth, $h = 1$. White solid lines indicate 95% tolerance contours.

between risk and distance from the incinerator, using a parametric model for the decay

in risk with increasing distance. The same conclusions were found by Gatrell (1990) who analysed the full set of data using the same methodology. A more extensive investigation was carried out by Elliott et al. (1992b). They analysed the incidence of cancers of the larynx and lung around the incinerator site in Lancashire and another nine sites of similar incinerators in Great Britain. Using a non-parametric method due to Stone (1988), they found no evidence of a consistent excess risk of either cancer among residents living close to the incinerators.

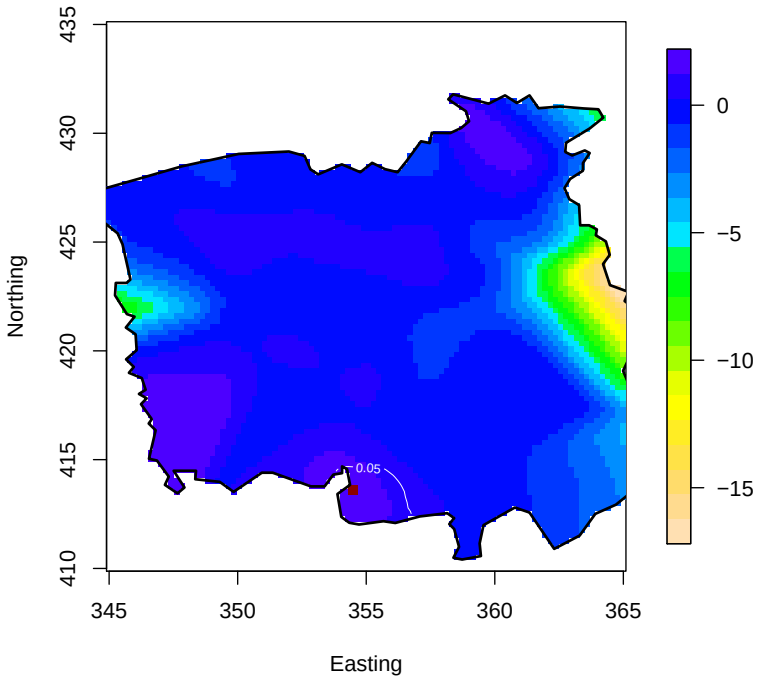

Figure 2.8: Estimates of the log-relative risk of larynx cancer in the Chorley-Ribble region of Lancashire, England. The estimate is computed using the density ratio method with subjective bandwidth $h = 1$. The dashed lines indicate 95% tolerance contours for areas of elevated risk. The red square represents the incinerator.

We computed the estimate of the log relative risk function using a bivariate Gaussian kernel with subjective bandwidth $h = 1$. Then we calculated 95% tolerance contours based on a p-value surface of 0.05 using an asymptotic z-test statistic. The P-values surface based on the asymptotic theory can be seen in the Figure 2.7. The estimated log relative risk surface with tolerance contours are depicted in Figure 2.8.

The suggestion of a small area of elevated risk in the south, as given by the p-value surface in Figure 2.7 and also the density ratio approach appeared in Figure 2.8, is in line with the aforementioned conjecture of the incinerator as a point source of risk.

2.6.2 Myrtle tree data

The second application related to the disease Myrtle Wilt for trees growing in Tasmania, Australia. Myrtle Wilt is a fungal disease of Myrtle beech (*Nothofagus cunninghami*), and is a major cause of mortality in this type of tree in the states of Tasmania and Victoria. The fungus is spread by air and water-borne spores, entering a new tree via wounds in its outer bark. It may also spread via root grafts or root contact in neighbouring trees. See for example, Kile, Packham and Elliott (1989) for further discussion.

Figure 2.9 displays a pattern of 106 diseased and 221 healthy Myrtle Beech trees in a 170.5×213.0 metre rectangular region. We have added an arbitrary polygonal window to this plot over which we compute our estimates of the relative risk of disease since it is useless to extend the estimates into the empty regions beyond this window.

As before, log-relative risk estimates (along with associated tolerance contours) are

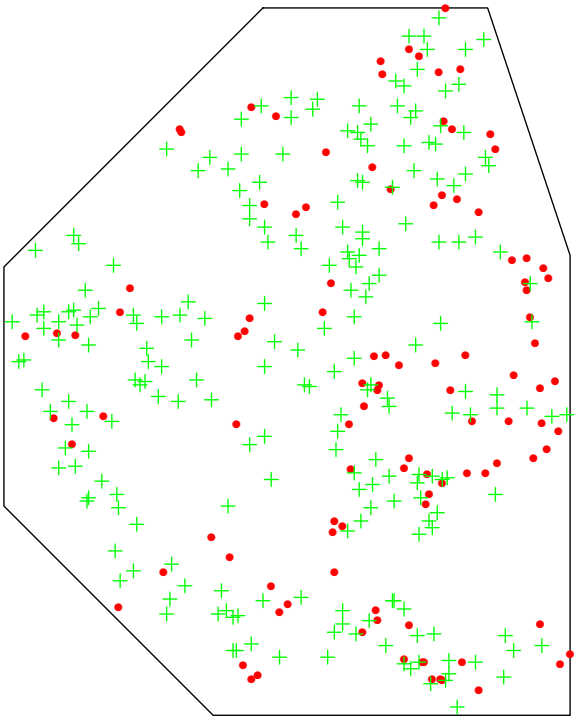

Figure 2.9: Plot of 106 diseased (•) and 221 healthy (+) Myrtle Beech trees in Tasmania.

computed on data using the density ratio method. A subjective bandwidth $h = 30$ is used. The resulting estimates are displayed in Figure 2.10, along with 95% tolerance contours. Eyeballing the data suggest that there is a stronger trend across the region from left to right.

2.7 Conclusion

In this Chapter, we provide a description of the spatial relative risk function in geographical epidemiology. As we understand, the most important factor in kernel

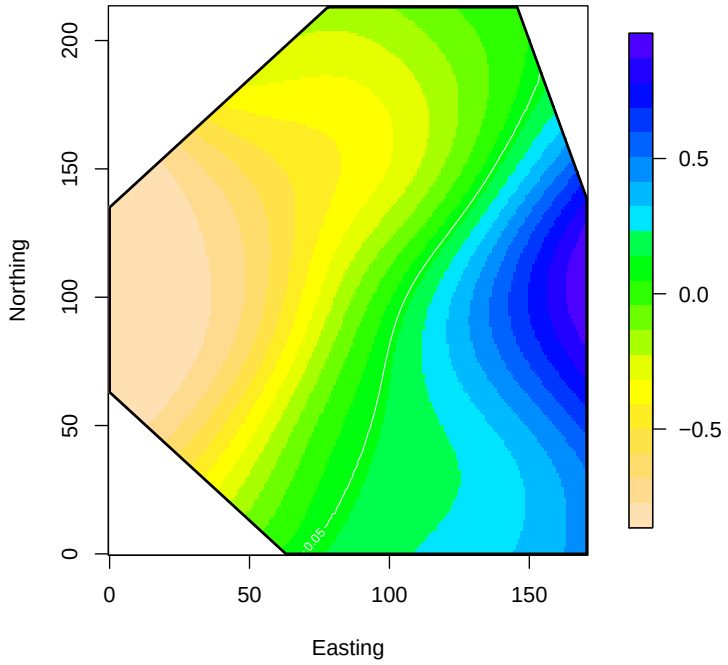

Figure 2.10: Estimates of the log-relative risk of disease from the Myrtle Beech data. This estimate is obtained by using the density ratio method with subjective bandwidth $h = 30$. The solid line indicates 95% tolerance contours for areas of elevated risk.

density estimation (and hence relative risk estimation) is the degree of smoothing. We do not give much attention to smoothing methods here, leaving this until Chapter 3. However, we use subjective bandwidths in this Chapter in the computation of density estimates.

We discussed two problems arising in the estimation of kernel density estimates and relative risk estimates. One of these is known as the edge effect, which is affected by

the observations occurred at the boundary of the region and the lack of information outside the region. The other difficulty is sparsely populated areas and hence makes relative risk estimates unstable.

The bias and variance of asymptotic properties of kernel estimates and hence relative risk estimates are obtained over a particular geographical region. We highlight relative risk hot-spots using tolerance contours based on a pointwise p-value surface. The use of kernel density and hence relative risk estimates is illustrated in two real data sets, Chorley-Ribble and Myrtle tree.

Chapter 3

Bandwidth selection for the density ratio estimator

3.1 Introduction

The practical realization of the kernel density estimator requires the specification of the bandwidth and so does density ratio estimator of relative risk. A bandwidth can be in the form of either a constant (h) or a matrix ($\mathbf{H}$). We focus on the former choice with previous work on relative risk estimation. The choice of bandwidth is very important since it significantly effects the kernel density estimator and hence the relative risk estimator. In density estimation, there have been many methods to choose the smoothing parameter. The simplest approach is to choose the bandwidth subjectively by eye. So one can begin with a bigger bandwidth and to decrease the degree of smoothing until one achieves a visually suitable level of smoothing. When the user has some knowledge about the structure of the data, this approach is more feasible. One disadvantage in this method is that it can be very time consuming to select the bandwidth by eye. Moreover, such a subjective approach is unlikely to be widely accepted amongst epidemiologists using the resultant relative risk estimates.

So we need to try other methods as well.

In choosing the smoothing parameter, Kelsall & Diggle (1995a) have carried out a simulation study for the univariate setting to compare the performance of different cross-validation bandwidth selectors. Their results suggest that one particular method labelled as "new cross-validation" is the preferred method for choice of smoothing parameter when estimating the log relative risk function, $\rho(\mathbf{x})$. After that they extend it to the two-dimensional case in Kelsall & Diggle (1995b). Hazelton & Davies (2009) discuss about the fixed bandwidth in kernel density estimates of the relative risk function. Davies & Hazelton (2010) work on the use of a spatially adaptive, variable smoothing parameter in relative risk estimation. Duong & Hazelton (2003) have worked on bandwidth matrices for bivariate kernel density estimation. They have developed a new type of plug-in selector for full bandwidth matrices.

This Chapter focuses on data-based bandwidth selections in bivariate kernel density estimation of the relative risk function. This is accomplished through numerical calculation of optimal bandwidths and error criteria, as well as simulation.

The rest of the sections are organized as follows. We discuss the mean integrated squared error (MISE), mean integrated absolute error (MIAE) and the asymptotic mean integrated squared error (AMISE) in Section 3.2. The optimal bandwidth formulaes using data-driven bandwidth selectors are obtained in Sections 3.3.1 and 3.3.2. The simulation study to compare such bandwidths is performed in Section 3.4 with the control densities and relative risk functions given in Table 3.1. These

bandwidth selectors are illustrated in the real data set, given in Section 3.5.

3.2 Error criteria

The choice of bandwidth plays a key role in determining the performance of relative risk estimators. We will seek bandwidth selectors that produce the optimal performance. Such performance of bandwidth selectors can be quantified by reducing the difference between a relative risk estimate, $\hat{r}$ (or $\hat{\rho}$ on log scale) with the true function, r (or ρ on log scale). We use existing error criteria to measure the difference. One well-known error criterion is the Integrated Squared Error (ISE). The ISE of $\hat{\rho}$ can be formulated as follows:

$$ISE\{\hat{\rho}(.;h)\} = \int_{\mathcal{R}} [\hat{\rho}(\mathbf{x};h) - \rho(\mathbf{x})]^2 d\mathbf{x}.$$

Here h $(= h_1 = h_2)$ is assumed a common bandwidth. The ISE is a random variable so an alternative is the Mean Integrated Squared Error or MISE, defined as follows:

$$MISE(h) = MISE\{\hat{\rho}(.;h)\} = \mathbb{E}\int [\hat{\rho}(\mathbf{x};h) - \rho(\mathbf{x})]^2 d\mathbf{x}.$$

Kelsall & Diggle (1995a) have considered the ISE and the MISE in the estimation of relative risk function (on log scale). Wand & Jones (1995) have mostly used the MISE for measuring the global performance of the kernel density estimator. In the context of density estimation, Devroye and Györfi (1985) have worked on Mean Integrated Absolute Error (MIAE), which is defined as follows:

$$MIAE\{\hat{\rho}(.;h)\} = \mathbb{E}\int |\hat{\rho}(\mathbf{x};h) - \rho(\mathbf{x})| d\mathbf{x}$$

However, we use the MISE since it is the most mathematically tractable criterion and is the most widely used in practice. So we wish to find the optimal bandwidth by

minimizing the MISE. That is,

$$h_{MISE} = \arg\min_h \text{MISE}[\hat{\rho}(;h)].$$

In the context of density estimation, Wand & Jones (1995) have mentioned that MISE does not have a closed form. So computing h_{MISE} in general is very challenging, but one can find an approximation to the MISE. We have

$$\begin{aligned}
\text{MISE} \;\; \hat{\rho}(\mathbf{x};h) &= \int_{\mathcal{R}} MSE \;\; \hat{\rho}(\mathbf{x};h)d\mathbf{x} \\
&= \int_{\mathcal{R}} \mathbb{E}[\hat{\rho}(\mathbf{x};h) - \rho(\mathbf{x})]^2 d\mathbf{x} \\
&= \int_{\mathcal{R}} [\mathbb{E}\hat{\rho}(\mathbf{x};h) - \rho(\mathbf{x})]^2 d\mathbf{x} + \int \text{Var}\hat{\rho}(\mathbf{x};h)d\mathbf{x} \\
&= \int_{\mathcal{R}} [\text{Bias}\hat{\rho}(\mathbf{x};h)]^2 d\mathbf{x} + \int \text{Var}\hat{\rho}(\mathbf{x};h)d\mathbf{x}
\end{aligned}$$

This gives the MISE as the sum of the integrated square bias and the integrated variance.

By substituting the asymptotic properties of bias and variance of $\hat{\rho}$, obtained in Chapter 2, we get

$$\begin{aligned}
\text{MISE}(h) &= \frac{h^4\mu_2(K)^2}{4}\int_{\mathcal{R}}\left[\frac{\nabla^2 f(\mathbf{x})}{f(\mathbf{x})} - \frac{\nabla^2 g(\mathbf{x})}{g(\mathbf{x})}\right]^2 d\mathbf{x} + \frac{R(K)}{h^2}\int_{\mathcal{R}}\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right] d\mathbf{x} \\
&\quad + o(h^4) + o(n_1^{-1}h^{-2} + n_2^{-1}h^{-2}).
\end{aligned} \tag{3.2.1}$$

We differentiate this MISE with respect to h and then set the derivative equal to zero. Then we reach to the optimal bandwidth h_{opt}.

$$h_{\text{opt}} = \left[\frac{2R(K)\displaystyle\int_{\mathcal{R}}\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right] d\mathbf{x}}{\mu_2(k)^2\displaystyle\int_{\mathcal{R}}\left[\frac{\nabla^2 f(\mathbf{x})}{f(\mathbf{x})} - \frac{\nabla^2 g(\mathbf{x})}{g(\mathbf{x})}\right]^2 d\mathbf{x}}\right]^{1/6}.$$

Note that we need to take care when estimating f and g since the denominator of h_{opt} disappears if $f = g$, which makes the optimal bandwidth large. The approximately optimal bandwidth, h_{opt} converges to zero as n_1 and/or n_2 increases. Furthermore, the approximate value of this bandwidth depends on the unknown densities being estimated. However, there are existing several methods for choosing h as discussed now.

3.3 Cross validation bandwidth selectors

3.3.1 Least squares cross validation

The Least Squares Cross Validation (LSCV) criterion is a fully automatic method for selecting the smoothing parameter. This technique was suggested by Rudemo (1982) and Bowman (1984) in the context of density estimation. See also Bowman, Hall and Titterington (1984), Hall (1983) and Stone (1984) for further reference. In this section, we discuss the process of computing LSCV estimator in the estimation of log relative risk, $\hat{\rho}$.

Suppose that f and g are probability density functions of cases and controls respectively defined within a given region $\mathcal{R}$. In order to select h for the estimation of the function $\alpha(f, g)$ (where, $\alpha(f(\mathbf{x}), g(\mathbf{x})) = \rho(\mathbf{x})$), a straightforward method is to choose the values which minimize the integrated squared error (ISE) as follows:

$$\begin{aligned} ISE\{\alpha(\hat{f},\hat{g})\} &= \int\int [\alpha(\hat{f}(\mathbf{x}),\hat{g}(\mathbf{x})) - \alpha(f(\mathbf{x}),g(\mathbf{x}))]^2 d\mathbf{x}. \\ &= \int\int [\hat{\alpha}^2 - 2\hat{\alpha}\alpha + \alpha^2]. \end{aligned}$$

This is equivalent to minimizing

$$LSCV(h) = \int\int \alpha(\hat{f},\hat{g})^2 d\mathbf{x} - 2\int\int \alpha(\hat{f},\hat{g})\alpha(f,g)d\mathbf{x} \tag{3.3.1}$$

since the last term does not depend on h and so can be omitted.

A first-order Taylor approximation to $\alpha(f,g)$ gives,

$$\alpha(f,g) \simeq \alpha(\hat{f},\hat{g}) + (f-\hat{f})\frac{\partial\alpha}{\partial\hat{f}} + (g-\hat{g})\frac{\partial\alpha}{\partial\hat{g}}.$$

Substituting the above equation in (3.3.1), we get

$$LSCV(h) = \int\int \alpha(\hat{f},\hat{g})^2 d\mathbf{x} - 2\int\int \alpha(\hat{f},\hat{g})\left[\alpha(\hat{f},\hat{g}) + (f-\hat{f})\frac{\partial\alpha}{\partial\hat{f}} + (g-\hat{g})\frac{\partial\alpha}{\partial\hat{g}}\right] d\mathbf{x}.$$

$$\begin{aligned} LSCV(h) = &-\int\int \alpha(\hat{f},\hat{g})^2 d\mathbf{x} + 2\int\int A_\delta(\hat{f},\hat{g})\hat{f}(\mathbf{x})d\mathbf{x} \quad + 2\int\int A_\tau(\hat{f},\hat{g})\hat{g}(\mathbf{x})d\mathbf{x} \\ &- 2\int\int A_\delta(\hat{f},\hat{g})f(\mathbf{x})d\mathbf{x} - 2\int\int A_\tau(\hat{f},\hat{g})g(\mathbf{x})d\mathbf{x} \end{aligned} \tag{3.3.2}$$

where $A_\delta(c_1,c_2) = \alpha(c_1,c_2)\dfrac{\partial\alpha(c_1,c_2)}{\partial c_1}$ and similarly $A_\tau(c_1,c_2) = \alpha(c_1,c_2)\dfrac{\partial\alpha(c_1,c_2)}{\partial c_2}$.

The last two terms in the above equation are expectations with respect to the unknown densities f and g. Therefore, we use the well-known "leave-one-out" criteria to estimate the expectations and hence the last two terms are replaced by

$$-2n_1^{-1}\sum_{i=1}^{n_1} A_\delta\{\hat{f}^{-i}(\mathbf{x}_i),\hat{g}(\mathbf{x}_i)\} - 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} A_\tau\{\hat{f}(\mathbf{x}_j),\hat{g}^{-j}(\mathbf{x}_j)\},$$

where

$$\hat{f}^{-i} = (n_1 - 1)^{-1} \sum_{j=1; j\neq i}^{n_1} K_h(\mathbf{x} - \mathbf{x}_j),$$

a density estimate of f constructed from all the data points except $(\mathbf{x}_i)$, and similarly

$$\hat{g}^{-j} = (n_2 - 1)^{-1} \sum_{i=n_1+1; i\neq j}^{n_1+n_2} K_h(\mathbf{x} - \mathbf{x}_i),$$

where n_1 and n_2 are case and control sample sizes respectively. Then equation (3.3.2) reduces to

$$\begin{aligned}
\widehat{LSCV}(h) &= -\int\int \alpha(\hat{f},\hat{g})^2 d\mathbf{x} + 2\int\int A_\delta(\hat{f},\hat{g})\hat{f}(\mathbf{x})d\mathbf{x} \\
&\quad + 2\int\int A_\tau(\hat{f},\hat{g})\hat{g}(\mathbf{x})d\mathbf{x} - 2n_1^{-1}\sum_{i=1}^{n_1} A_\delta(\hat{f}^{-i},\hat{g}) - 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} A_\tau(\hat{f},\hat{g}^{-j}) \\
&= -\int\int \alpha(\hat{f},\hat{g})^2 d\mathbf{x} + 2\int\int \alpha(\hat{f},\hat{g})\hat{f}\left(\frac{1}{\hat{f}}\right) d\mathbf{x} + 2\int\int \alpha(\hat{f},\hat{g})\hat{g}\left(\frac{-1}{\hat{g}}\right) d\mathbf{x} - \\
&\quad 2n_1^{-1}\sum_{i=1}^{n_1} \alpha(\hat{f}^{-i},\hat{g})\left(\frac{1}{\hat{f}^{-i}}\right) - 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} \alpha(\hat{f},\hat{g}^{-j})\left(\frac{-1}{\hat{g}^{-j}}\right) \\
&= -\int\int \alpha(\hat{f},\hat{g})^2 d\mathbf{x} - 2n_1^{-1}\sum_{i=1}^{n_1} \alpha(\hat{f}^{-i},\hat{g})\{\hat{f}^{-i}\}^{-1} + 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} \alpha(\hat{f},\hat{g}^{-j})\{\hat{g}^{-j}\}^{-1}
\end{aligned}$$

By replacing $\alpha(f,g)$ with $\log(f/g) = \rho$, we get

$$\begin{aligned}
\widehat{LSCV}(h) &= -\int\int \{\hat{\rho}_h(\mathbf{x})\}^2 d\mathbf{x} - 2n_1^{-1}\sum_{i=1}^{n_1} \log\left[\frac{(\hat{f}_h^{-i}(\mathbf{x}_i))}{\hat{g}_h(\mathbf{x}_i)}\right]\{\hat{f}_h^{-i}(\mathbf{x}_i)\}^{-1} + \\
&\quad 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} \log\left[\frac{(\hat{f}_h(\mathbf{x}_j))}{\hat{g}_h^{-j}(\mathbf{x}_j)}\right]\{\hat{g}_h^{-j}(\mathbf{x}_j)\}^{-1}. \qquad (3.3.3)
\end{aligned}$$

When $\mathbf{x}$ lies at the boundary of the region, as discussed in Chapter 2, the edge correction technique needs to be applied. However, since we use a common bandwidth

h for the computation of $\hat{f}$, $\hat{g}$ and then the equation (3.3.3) is revised with the edge correction factors as follows:

$$\begin{aligned}\widehat{LSCV}(h) &= -\int\int\{\hat{\rho}_h(\mathbf{x})\}^2 d\mathbf{x} - 2n_1^{-1}\sum_{i=1}^{n_1}\log\left[\frac{(\hat{f}_h^{-i}(\mathbf{x}_i))}{\hat{g}_h(\mathbf{x}_i)}\right]\{\tilde{f}_h^{-i}(\mathbf{x}_i)\}^{-1} + \\ &\quad 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2}\log\left[\frac{(\hat{f}_h(\mathbf{x}_j))}{\hat{g}_h^{-j}(\mathbf{x}_j)}\right]\{\tilde{g}_h^{-j}(\mathbf{x}_j)\}^{-1}, \qquad (3.3.4)\end{aligned}$$

where,

$$\tilde{f}^{-i}(\mathbf{x}) = (n_1-1)^{-1}\sum_{j=1;j\neq i}^{n_1} K_h(\mathbf{x}-\mathbf{x}_j)/q_1(\mathbf{x}).$$

Similarly,

$$\tilde{g}^{-j}(\mathbf{x}) = (n_2-1)^{-1}\sum_{i=n_1+1;i\neq j}^{n_1+n_2} K_h(x-x_j)/q_2(\mathbf{x}),$$

where $q_1(\mathbf{x})$ and $q_2(\mathbf{x})$ are the edge correction terms, defined in Chapter 2. The minimization of the above estimate $\widehat{LSCV}$ over h gives LSCV optimal bandwidth, namely,

$$\hat{h}_{LSCV} = \arg\min_h \widehat{LSCV}(h).$$

In practice, the LSCV bandwidth selector has drawbacks. It often has more than one local minimum with some spurious ones often leading to under smoothing. Hall & Marron (1991a) have worked on this problem in the context of density estimation. Simulation studies have shown that this problem can often be well handled by selecting the largest value of h for which a local minimum occurs. It is well known that LSCV tends to produce highly variable bandwidth selectors in many problems. See Hall & Marron (1987a), Park & Marron (1990) for further details. We therefore may expect it to do so when estimating ρ. We examine this further through our numerical studies.

3.3.2 Likelihood cross validation

Likelihood cross validation (LCV) method is motivated by the maximum likelihood principle. See Stone (1974) and Geisser (1975). Clark & Lawson (2004) have studied different bandwidths including LCV in the context of a binary regression problem. They note that the conditional probability of a case, for a given point $\mathbf{x}$ is

$$\begin{aligned} p(\mathbf{x}) &= \Pr\{Y_i = 1 | \mathbf{X} = \mathbf{x}_i\} \\ &= \frac{n_1 f(\mathbf{x})}{n_2 g(\mathbf{x}) + n_1 f(\mathbf{x})}, \end{aligned}$$

where Y_i are the binary labels giving one if the i^{th} point is a case and zero if it is a control. This is directly related to relative risk, $r(\mathbf{x})$ by $p(\mathbf{x}) = n_1 r(\mathbf{x})/(n_2 + n_1 r(\mathbf{x}))$. The conditional log likelihood is

$$\log L\{\hat{p}(\mathbf{x})\} = \log \left[\prod_{i=1}^{n} \{\hat{p}(\mathbf{x}_i; h)\}^{y_i} \{1 - \hat{p}(\mathbf{x}_i; h)\}^{1-y_i} \right].$$

In the non-parametric case, this maximum likelihood principle would lead to unsatisfactory estimates $p(\mathbf{z}_i) = 0$ or $p(\mathbf{z}_i) = 1$ according to the binary variables $y_i = 0$ or $y_i = 1$ respectively. According to the LCV criteria, we consider the log likelihood leave-one-out function that is minimized with respect to h, taken as its estimator.

That is,

$$\begin{aligned}
\widehat{LCV}(h) &= \log\left[\prod_{i=1}^{n}\{\hat{p}^{-i}(\mathbf{x}_i;h)\}^{y_i}\{1-\hat{p}^{-i}(\mathbf{x}_i;h)\}^{1-y_i}\right] \\
&= \sum_{i=1}^{n}\left[y_i\log[\hat{p}^{-i}(\mathbf{x}_i;h)]+(1-y_i)\log[1-\hat{p}^{-i}(\mathbf{x}_i;h)]\right] \\
&= \sum_{i=1}^{n_1}\left[y_i\log[\hat{p}_1^{-i}(\mathbf{x}_i;h)]+(1-y_i)\log[1-\hat{p}_1^{-i}(\mathbf{x}_i;h)]\right]+ \\
&\qquad \sum_{i=n_1+1}^{n_1+n_2}\left[y_i\log[\hat{p}_2^{-i}(\mathbf{x}_i;h)]+(1-y_i)\log[1-\hat{p}_2^{-i}(\mathbf{x}_i;h)]\right] \\
&= \sum_{i=1}^{n_1}\log[\hat{p}_1^{-i}(\mathbf{x}_i;h)]+\sum_{i=n_1+1}^{n_1+n_2}\log[1-\hat{p}_2^{-i}(\mathbf{x}_i;h)]
\end{aligned}$$

where

$$\hat{p}_1^{-i}(\mathbf{x}_i;h)=\frac{n_1\hat{f}^{-i}(\mathbf{x}_i,h)}{n_2\hat{g}(\mathbf{x}_i;h)+n_1\hat{f}^{-i}(\mathbf{x}_i,h)},$$

$$\hat{p}_2^{-i}(\mathbf{x}_i;h)=\frac{n_1\hat{f}(\mathbf{x}_i;h)}{\left[n_1\hat{f}(\mathbf{x}_i;h)+n_2\hat{g}^{-i}(\mathbf{x}_i;h)\right]}$$

and $n=n_1+n_2$.

Also $\hat{f}^{-i}$ and $\hat{g}^{-i}$ represent the kernel density estimates constructed respectively from all case and control data points except the i^{th} datum. The above results are obtained when the points occur interior to the region. However, when the points fall at the boundary of the region, the above results remain as it is since the edge correction terms are cancelled out on top and bottom of $\hat{p}_1^{-i}$ and $\hat{p}_2^{-i}$ terms when using the same bandwidth h for cases and controls. The likelihood cross-validation choice of h is

$$\hat{h}_{\text{LCV}}=\arg\max_h\widehat{\text{LCV}}(h).$$

It has been noted by Scott & Factor (1981) in the context of density estimation that the performance of LCV bandwidth is very sensitive to outliers. It is a fact that LCV

does not directly attempt to minimize MISE($\hat{\rho}$). Indeed, this methodology is more natural when aiming to estimate $p(\mathbf{x})$ (as was the aim of Clark & Lawson, 2004) than when estimating $\rho(\mathbf{x})$.

3.4 Simulation study to compare LSCV over LCV in $\hat{\rho}$ estimation

In order to examine the practical performance of bandwidth selectors in the estimation of $\hat{\rho}$, we conduct a simulation study to compare LSCV over LCV.

We consider six synthetic problems, in each defined on a square region $\mathcal{R}$ =[0,10]×[0,10]. The forms of control densities, g and relative risk functions, r are displayed in Table 3.1. These problems are designed from real-world population distributions with a mixture of rural and urban areas. In particular, we consider control densities in a single mode because the results will also provide guidance to cases where multiple modes are very well separated as we might expect to see with city and towns with real applications. So samples of controls (n_2) are randomly generated from the defined control densities while sample of cases (n_1) are generated from the control distribution (g), then accepted/rejected with probabilities proportional to the relevant relative risk function (r) until the desired number n_1 is reached.

Each control density is defined to be either uniform or bivariate normal density (truncated to $\mathcal{R}$). The control densities g are displayed in Figure 3.1. The corresponding relative risk (on log scale) functions ρ are given in Figure 3.2 as filled contour plots.

Table 3.1: Control densities and relative risk functions for six synthetic models. The function ϕ_σ is a bivariate normal density with zero mean vector and covariance matrix $\sigma^2 I$, where I is the 2×2 identity matrix. In addition, $\phi_\sigma^{\mathcal{R}}$ denotes ϕ_σ truncated to $\mathcal{R}$. The location parameters are $\mu_1 = [4,4]^T$, $\mu_2 = [5,5]^T$ and $\mu_3 = [6,6]^T$.

Problem 1	$g(z) = 0.01$ $r(z) \propto 1 + \exp(-\|z - \mu_2\|^2)$
Problem 2	$g(z) = 0.01$ $r(z) \propto 1 + 4\ \exp(-\|z - \mu_2\|^2)$
Problem 3	$g(z) = 0.01$ $r(z) \propto 1 + 2\ \exp(-\|z - \mu_1\|^2) + 2\ \exp(-3\|z - \mu_3\|^2)$
Problem 4	$g(z) = \phi^R_{1/\sqrt{2}}(z - \mu_2)$ $r(z) \propto 1 + \exp(-\|z - \mu_2\|^2)$
Problem 5	$g(z) = \phi^R_{1/\sqrt{2}}(z - \mu_2)$ $r(z) \propto 1 + 4\ \exp(-\|z - \mu_2\|^2)$
Problem 6	$g(z) = \phi^R_{1/\sqrt{2}}(z - \mu_2)$ $r(z) \propto 1 + 2\ \exp(-\|z - \mu_1\|^2) + 2\ \exp(-3\|z - \mu_3\|^2)$

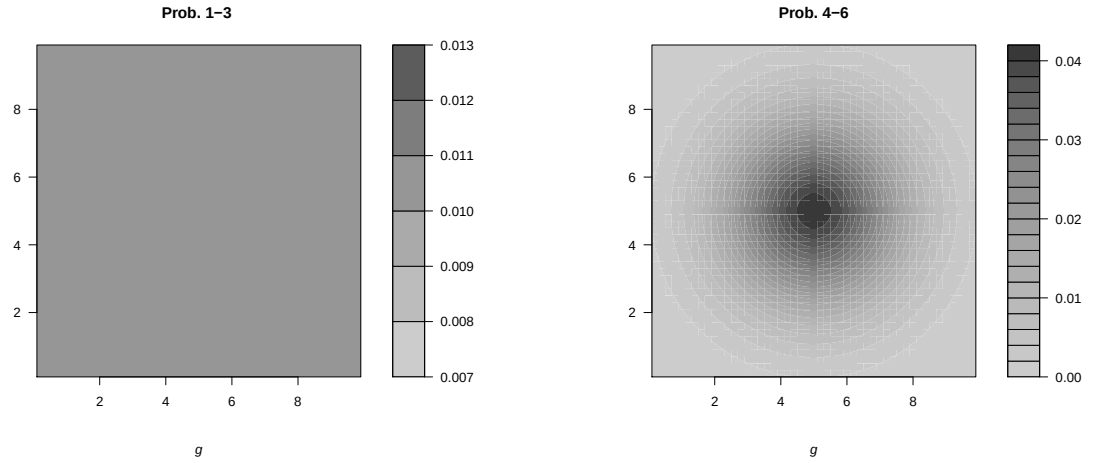

Figure 3.1: Filled contour plots of the control densities, uniform(left panel) and bivariate normal (right panel) as described in the text.

There are three relative risk functions used. The first relative risk function has a mild circular clustering of risk in the middle of the region, the second consists a strong circular clustering at the centre and the last function has an average risk variation with two hot spots.

For each Problem, we carry out three sets of sample sizes. In the first set, we use 100 cases and controls ($n_1 = n_2 = 100$); in the second set, 100 cases and 400 controls ($n_1 = 100$ $n_2 = 400$) and in the third set, 400 cases and controls ($n_1 = n_2 = 400$). For each set, we generate 400 case-control data sets. The log relative risk function is computed for each case using the density ratio method. We use LSCV and LCV bandwidth selectors, which are obtained through an optimization process. Due to multiple local minima of LSCV estimator, we restrict the search range to the interval $(h_0, 4h_0)$, where h_0 is the geometric mean of the normal reference bandwidths (for

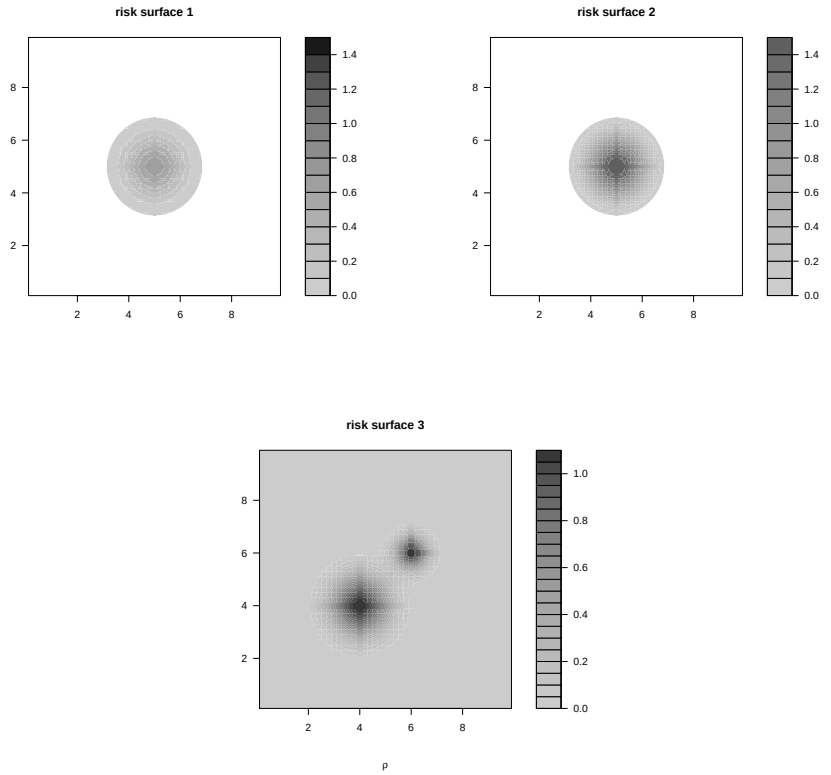

Figure 3.2: Filled contour plots of the log-relative risk functions as described in Table 3.1. Problems 1 and 4 represent the risk surface 1. Problems 2 and 5 represent the risk surface 2. Problems 3 and 6 represent the risk surface 3.

e.g. Wand & Jones, 1995) for case and control density estimates. Bandwidths are optimal if the targeted densities are normal. To assess the performance, we compute the integrated squared error (ISE) of each estimate $\hat{\rho}$ over the region $\mathcal{R}$. Log(ISE) are displayed using box-and-whisker plots which split by the bandwidth type (either LSCV or LCV) and set in Figure 3.3. These all computations were obtained using R statistical software. See R Development Core Team (2010).

Overall the simulation results shows inferior performance of the LCV bandwidth selector in density ratio estimation. LSCV bandwidth selector performs better. Specifically, Problems 5 and 6 have particularly significant differences between these two bandwidth selectors. The log(ISE) for Problems 1, 2, 3 and 4 is more comparable between LSCV and LCV methods. For both bandwidth selectors, it is vital to note that the selected value of h occurred at a boundary of the search interval $[h_0, 4h_0]$ for some simulated data sets. This reveals problems with the variability of the performance of LCV and LSCV. There may be two possible reasons for the significant differences of the bandwidth selectors for Problems 5 and 6. Perhaps because these Problems are the most challenging in terms of the complexity of ρ or because they require the smallest values of h.

The LSCV approach works by minimizing an approximate estimator of MISE (modulo irrelevant additive constants). Usually, this method suffers from multiple local minima and high variability. The LCV method optimizes a leave-one-out estimate of the log-likelihood. The drawback of LCV is that it is sensitive to outliers and hence can create an over-smoothed estimate. However, neither method is perfect. According to the above simulation study, LSCV performs better than LCV as the former explicitly targets the MISE of $\hat{\rho}$ while the latter does not.

3.5 Real application: Cancers in South Lancashire

The distribution of the data in this example is already displayed in Chapter 1 and the subject of a preliminary analysis in Chapter 2. See Diggle (1990) for further details.

We estimate log relative risk functions using density ratio method for Chorley data. Bandwidths 0.78 and 2.74 were obtained using the LSCV and LCV respectively. The resulting estimates are depicted in Figure 3.4. The bandwidth estimators provide different interpretations. The LSCV bandwidth allows the relative risk estimator to highlight the areas having elevated risk (95% tolerance contours) around the incinerator. In the mean time, LCV gives a smoothed risk surface, having high risk in the southwest.

3.6 Conclusion

The choice of bandwidth plays a key role in determining the performance of kernel density estimators and hence relative risk estimators. This is a vital problem both from a practical and theoretical perspective. The contributions in this Chapter have been to review some existing methods, compare LCV against LSCV through a simulation, organize the asymptotic properties and look at the performance of these estimators through a real application.

According to these finite sample performance results, LSCV bandwidth selector seems at present to be the best option when implementing the density ratio estimator, but we note that this is largely a reflection of the poor performance of the alternative method. Also the simulation results depends on the chosen densities. We believe that the absence of an effective plug-in bandwidth selector for the density ratio method constitutes a major challenge for relative risk estimation in spatial epidemiology.

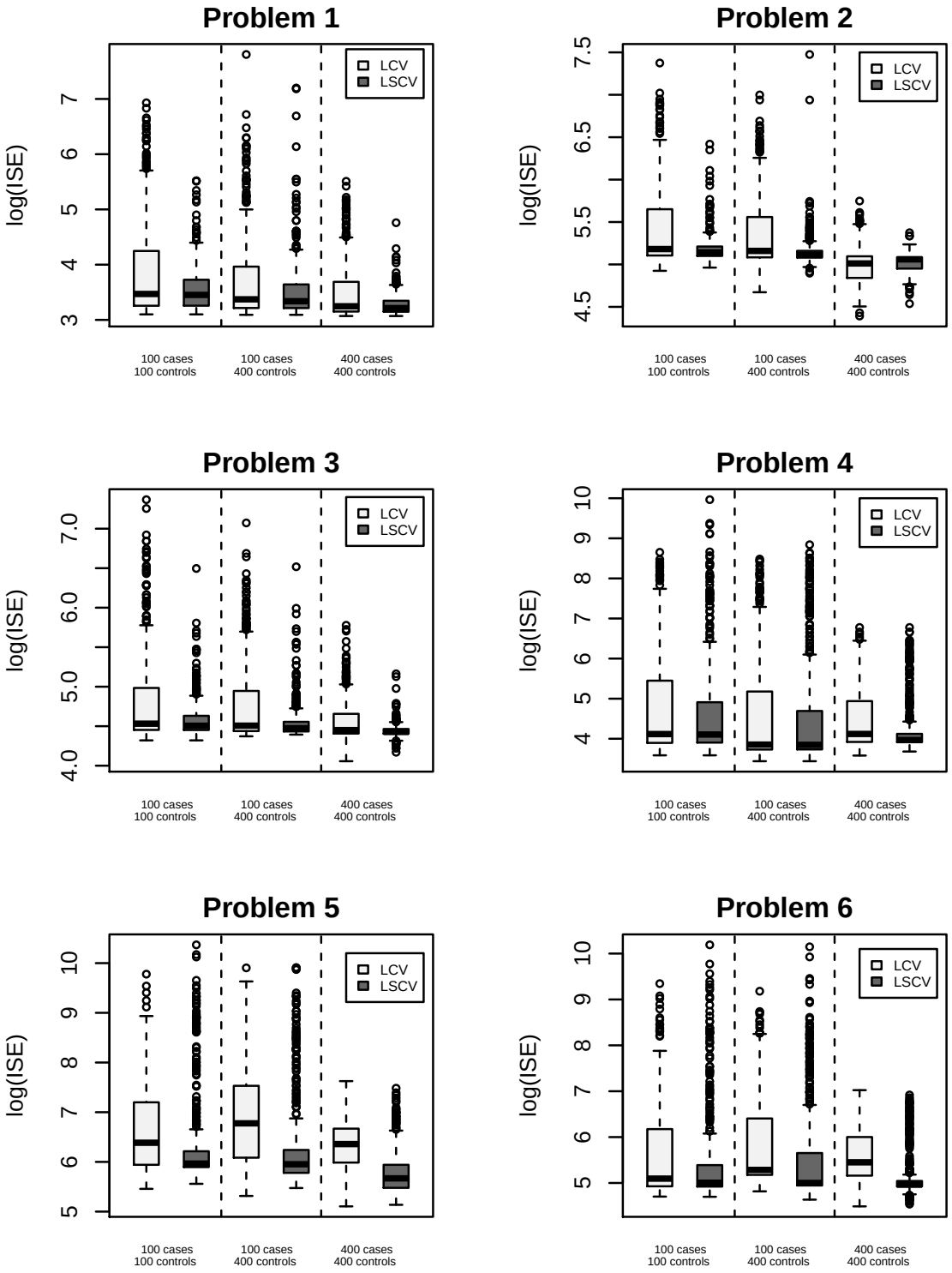

Figure 3.3: Boxplots of log(ISE) of log-relative risk estimates for problems 1-6 from Table 3.1. LCV and LSCV stand for likelihood and least squares cross validation bandwidths respectively.

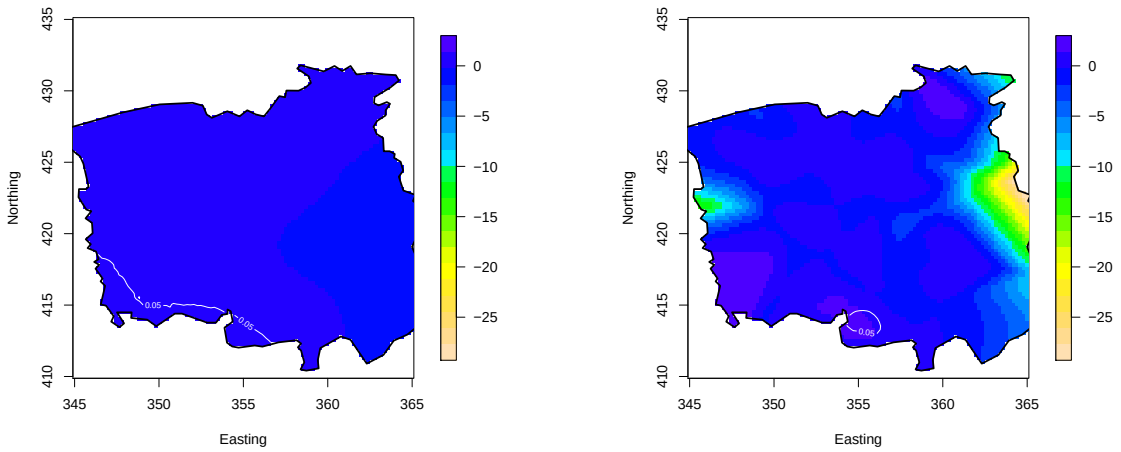

Figure 3.4: Estimates of log-relative risk of larynx cancer data. Left panel shows the estimate using LCV bw (2.74) while the right panel, using the LSCV bw (0.78).

Chapter 4

Local linear estimation of the relative risk function

4.1 Introduction

As we have seen, the relative risk function can be estimated by the ratio of kernel density estimates constructed from case and control location data respectively. This methodology has been used successfully in a variety of applications in both human and veterinary epidemiology (e.g. Sabel et al., 2000; Prince et al., 2001; Berke, 2005; Wheeler, 2007). In particular, this direct density ratio estimator of the relative risk function has proven adept at identifying approximately circular or elliptical areas of elevated risk, perhaps around a postulated point sources of risk. However, this approach can be less natural when the relative risk has a more global trend, as one might expect to see when there is a line source of risk such as a polluted river or a road.

An alternative approach to estimate the log-relative risk function is local polynomial regression. The idea of local polynomial regression has been around for a long time. It was systematically studied by Stone (1977, 1980, 1982) and Cleveland (1979). The work on local polynomial fitting includes Fan (1992, 1993), Fan & Gijbels (1992) and Ruppert & Wand (1994) in the context of standard regression. These papers give a detailed picture of the advantages of local polynomial fitting. These estimate the regression function at a particular point by locally fitting a p^{th} degree polynomial to the data.

One of the most important choices is the degree of the polynomial. Boundary bias considerations indicate that one should fit local lines. But for estimating regions of high curvature of the true function, such as at peaks and valleys, local line estimators can have a substantial bias. This problem can be avoided by fitting higher degree polynomials, although there are costs in terms of increased variance and computational complexity which needs to be considered. See Fan & Gijbels (1996).

As a special case, the Nadaraya-Watson estimator corresponds to fitting degree zero polynomials, which is locally constant. See Nadaraya (1964) and Watson (1964). Of particular importance and simplicity is the local linear estimator corresponding to $p = 1$.

Throughout this chapter, we focus on the degree one polynomial ($p = 1$ case), i.e. local linear regression. Local linear fitting has been seen to have several fascinating qualities in terms of intuitive and mathematical simplicity. One noteworthy feature is

the better performance at the boundaries compared to the local constant estimators (i.e. $p = 0$ case).

Local polynomial regression has been used for non-normal data. In particular, Clark & Lawson (2004) use local linear method to estimate the conditional probability function, p in disease mapping, using case-control data. However, the use of local linear regression for estimation of the (log) relative risk functions *per se* has not been examined in the literature. Therefore, what we have done in this chapter becomes a novel contribution to the relevant field.

This Chapter is divided into the following Sections. Firstly, we provide a detailed account of local linear regression in the estimation of the log-relative risk function in Section 4.2. During this estimation process, we use a local scoring algorithm to fit the local linear model in Section 4.3. Then we provide an analysis of asymptotic properties of local linear estimator. After that we construct tolerance contours to highlight the regions which show significantly elevated risk. We also examine data-driven bandwidth selection methods for the local-linear estimator, including a novel plug-in methodology in Section 4.5.1. As Kelsall & Diggle (1995a) noted, the use of plug-in bandwidth selectors in relative risk estimation is attractive in principle, since plug-in methods are generally preferable to (unsmoothed) cross-validation methods for kernel density estimation (e.g. Wand & Jones, 1995). However, Kelsall and Diggle (1998) noted that there are significant difficulties in implementing plug-in methods for the density ratio estimator because of the difficult forms of some of the functionals that require pilot estimation, and no further progress has been made in this direction.

We show here that plug-in bandwidth selection is a feasible prospect for local linear estimators.

In order to explore the practical performance of the density ratio estimator against the local linear estimator, we present the results of estimators by using the optimal smoothing in Section 4.6.1 and then by using a data-driven bandwidth in Section 4.6.2. In the latter, we present numerical results comparing the performance of the density ratio and local linear estimators using a variety of data-driven bandwidth selectors. We include a simulation study involving 32 different test scenarios, characterized by factors including the 'shape' of the risk surface (e.g. largely linear versus circular contours); the magnitude of the maximum relative risk; and the rate at which the risk decays from point or line sources.

We also examine analysis of three real data sets, the first relating to the distribution of Myrtle Wilt in a region of forest in the Australian state of Tasmania, the second concerned with the incidence of cancer of the larynx in the Chorley-Ribble region of Lancashire, England and the last data set describes about the foot and mouth epidemic (FMD) in 2001 of Cumbria, in the north-west of England. We computed relative risk estimates (on log scale) for both methods and constructed tolerance contours to highlight the regions of significantly high risk.

4.2 Local linear estimator of the relative risk function

Suppose that we need to estimate the log relative risk function at the point $\mathbf{x}$. In this section, we discuss how the estimation process of log relative risk in geographical epidemiology is done using local polynomial approach with the order p first, and then the local linear case as a novel work. In the local polynomial approach, we model log relative risk (ρ) in the neighborhood of $\mathbf{x}$ as a polynomial ; i.e.

$$\rho(\mathbf{z}) \approx P(\mathbf{z} - \mathbf{x}) \tag{4.2.1}$$

where P is a polynomial form of order p. Since $\rho(\mathbf{x}) \approx P(\mathbf{0})$ according to this polynomial approximation it follows that we can estimate $\rho(\mathbf{x})$ as the constant term from a suitably fitted polynomial $\breve{P}$.

In order to fit this model, we recast it as a binary regression. To do so, we introduce the binary variable Y, such that $Y_i = 1$ if $\mathbf{x}_i$ is a case location, and $Y_i = 0$ if $\mathbf{x}_i$ is a control location. Conditional on the sample sizes n_1 and n_2, we denote the probability that the observation i is a case rather than a control by

$$P(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) = p(\mathbf{x}_i).$$

Then the binary regression function p can be expressed as

$$\begin{aligned} p(\mathbf{x}) &= \frac{\pi f(\mathbf{x})}{(1-\pi)g(\mathbf{x}) + \pi f(\mathbf{x})} \\ &= \frac{\pi r(\mathbf{x})}{(1-\pi) + \pi r(\mathbf{x})} \end{aligned}$$

where $r(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})}$, is the relative risk function and $\pi = \frac{n_1}{n_1+n_2}$, is the overall proportion of cases. Then we get,

$$r(\mathbf{x}) = \frac{(1-\pi)}{\pi}\frac{p(\mathbf{x})}{(1-p(\mathbf{x}))}.$$

Following this on the logit scale in order to preserve the estimate in the range [0,1], we have

$$\begin{aligned}\log\left\{\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right\} &= \log\{r(\mathbf{x})\} + \log\left[\frac{n_1}{n_2}\right] \\ &= \rho(\mathbf{x}) + \log\left[\frac{n_1}{n_2}\right].\end{aligned}$$

Applying the local polynomial approximation for ρ from (4.2.1), therefore leads to a local polynomial logistic regression model:

$$\log\left\{\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right\} \approx Q(\mathbf{z}-\mathbf{x}) \tag{4.2.2}$$

where $P = Q - \log(n_1/n_2)$.

As noted by Signorini & Jones (2004) in the context of univariate binary regression, the natural way to fit the local polynomial representation in equation (4.2.2) is by maximizing a local likelihood of Tibshirani & Hastie (1987). If we define the local weights through a kernel function then the weighted log-likelihood can be written as,

$$L(Q,\mathbf{x}) = \sum_{i=1}^{n}\left(y_i Q(\mathbf{x}_i-\mathbf{x}) - \log\left\{1+\exp[Q(\mathbf{x}_i-\mathbf{x})]\right\}\right)K_h(\mathbf{x}_i-\mathbf{x}). \tag{4.2.3}$$

For estimation at $\mathbf{x} \in \mathcal{R}$, the fitted polynomial $\breve{Q}$ is the maximizer of L and hence the local polynomial regression estimator of the log-relative risk function is $\breve{\rho}(\mathbf{x}) = \breve{Q}(\mathbf{0}) - \log(n_1/n_2)$. In general the coefficients of $\breve{Q}$ will not be available in closed

form, so that estimation must proceed via numerical maximization of (4.2.3). The estimating equations requiring iterative solution are (Signorini & Jones, 2004),

$$\sum_{i=1}^{n}(\boldsymbol{Y}_i - \hat{p}(\mathbf{x}_i - \mathbf{x}))K_h(\mathbf{x}_i - \mathbf{x}) = 0 \tag{4.2.4}$$

where $n = n_1 + n_2$ and

$$\sum_{i=1}^{n}(\mathbf{Y}_i - \hat{p}(\mathbf{x}_i - \mathbf{x}))(\mathbf{x}_i - \mathbf{x})K_h(\mathbf{X}_i - \mathbf{x}) = 0$$

where $\text{logit}(\hat{p}(\mathbf{z})) = \hat{\beta}_0 + \hat{\beta}_1\mathbf{z}$ and that is,

$$\log\left(\frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})}\right) = \hat{\beta}_0 + \hat{\beta}_1\mathbf{z}.$$

The simplest form of our model is when the local polynomial is constant; i.e. $Q(\cdot) = \beta_0 \equiv \beta_{0,\mathbf{x}}$ (emphasizing the otherwise implicit dependence on the estimation point in the second case). When the local polynomial is constant,

$$\hat{p}(\mathbf{z}) = \frac{e^{\breve{\beta}_0}}{1+e^{\breve{\beta}_0}}$$

and the equation (4.2.4) becomes

$$\sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x}) = \sum_{i=1}^{n} \hat{p}(\mathbf{x}_i - \mathbf{x})\ K_h(\mathbf{x}_i - \mathbf{x})$$

By substituting $\hat{p}$, we get the maximizer $\breve{\beta}_0$ of the local likelihood equation satisfies

$$\sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x}) = \sum_{i=1}^{n} \frac{e^{\breve{\beta}_0}}{1+e^{\breve{\beta}_0}} K_h(\mathbf{x}_i - \mathbf{x}). \tag{4.2.5}$$

From equation (4.2.5), we get

$$\begin{aligned}
\breve{\beta}_0 &= \log\left[\frac{\sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x})}{\sum_{i=1}^{n} K_h(\mathbf{x}_i - \mathbf{x}) - \sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x})}\right] \\
&= \log\left[\sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x})\right] - \log\left[\sum_{i=1}^{n}(1-y_i)K_h(\mathbf{x}_i - \mathbf{x})\right].
\end{aligned}$$

We find that the local constant estimator of the log-relative risk, ρ_{LC}, is therefore given by

$$\begin{aligned}
\breve{\rho}_{\mathrm{LC}}(\boldsymbol{x}) &= \breve{\beta}_0 - \log(n_1/n_2) \\
&= \log\left[\sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x})\right] - \log\left[\sum_{i=1}^{n}(1-y_i)K_h(\mathbf{x}_i - \mathbf{x})\right] - \log[n_1/n_2] \\
&= \log\left[n_1^{-1}\sum_{i=1}^{n} y_i K_h(\mathbf{x}_i - \mathbf{x})\right] - \log\left[n_2^{-1}\sum_{i=1}^{n}(1-y_i)K_h(\mathbf{x}_i - \mathbf{x})\right] \\
&= \log\left[\hat{f}(\mathbf{x})\right] - \log\left[\hat{g}(\mathbf{x})\right], \qquad (4.2.6)
\end{aligned}$$

demonstrating the equivalence with the density ratio method, which is already described in Chapter 2.

In the local linear case we have

$$Q(\mathbf{z}) = \beta_0 + \beta_1(z_1 - x_1) + \beta_2(z_2 - x_2). \qquad (4.2.7)$$

The local linear estimator is then given by $\breve{\rho}_{LL}(\mathbf{x}) = \breve{\beta}_0 - \log(n_1/n_2)$ where $\breve{\beta}_0$ now denotes the first element of $\breve{\beta} = (\breve{\beta}_0, \breve{\beta}_1, \breve{\beta}_2)^T$, the polynomial coefficients maximizing (4.2.3) with Q as defined in (4.2.7). Local constant is a particular case of local linear, and so in theory the latter will be no worse than the former. However, in practice, the additional complexity of the local linear estimator can result in worse performance. As there is no explicit representation of $\breve{\rho}_{LL}(\mathbf{x})$, $\breve{\beta}$ will be fitted using an alternative method described in detail in the next section.

4.3 Local scoring procedure

Generalized additive models (GAM) were popularized as non-parametric regression techniques (Hastie and Tibshirani, 1990). They consist of assuming a parametric

Table 4.1: The local scoring procedure

---

1. Initialise $\hat{g}(\mathbf{x}_i; h) = 0$ and $\hat{\alpha} = \text{logit}\{n_2/(n_1 + n_2)\}$ (here $\hat{g}(\mathbf{x}_i; h) = \text{logit}\{p(\mathbf{x})\}$).

2. Set $\hat{\eta}_i = \hat{g}(\mathbf{x}_i; h)$ and $\hat{p}(\mathbf{x}_i) = \exp(\hat{\eta}_i)/\{1 + \exp(\hat{\eta}_i)\}$.

3. Construct the adjusted dependent variable $\mathbf{z}_i = \hat{\eta}_i + \dfrac{\mathbf{y}_i - \hat{p}(\mathbf{x}_i)}{\hat{p}(\mathbf{x}_i)\{1 - \hat{p}(\mathbf{x}_i)\}}$ with weights $w_i = \hat{p}(\mathbf{x}_i)\{1 - \hat{p}(\mathbf{x}_i)\}$

4. Fit a weighted linear model using either the Nadaraya-Watson regression estimator $\hat{g}(\mathbf{x}; h) = \dfrac{\sum_{i=1}^{n} w_i K_h(\mathbf{x} - \mathbf{x}_i)\mathbf{y}_i}{\sum_{i=1}^{n} w_i K_h(\mathbf{x} - \mathbf{x}_i)}$ or the local linear regression estimator.

5. Repeat step (2) replacing $\hat{\eta}_i^{(j)}$ by $\hat{\eta}_i^{(j+1)}$ until $\Delta(\hat{\eta}_i^{(j+1)}, \hat{\eta}_i^{(j)}) = \dfrac{\sum_{k=1}^{p} \|\hat{g}_k^{j+1}(\mathbf{x}; h) - \hat{g}_k^{j}(\mathbf{x}; h)\|}{\sum_{k=1}^{p} \|\hat{g}_k^{j}(\mathbf{x}; h)\|}$ is below some small threshold (e.g. 1E-5).

---

model for the data, but the relationship between the response variable and the non-response variable is non-parametric. For GAM, the backfitting iteration is performed by transforming the original observations to scores and fitting those in an iterative and weighted manner using backfitting. The overall procedure is called local scoring (Hastie and Tibshirani, 1990, p140), which generalizes the Fisher scoring procedure described in McCullagh and Nelder (1989, p.42).

Local scoring splits the multivariate regression into a sequence of univariate regressions that are easier to compute. Therefore, we use local scoring algorithm, displayed in Table (4.1) reproduced from Hastie and Tibshirani (1990) to fit the local linear model.

4.4 Asymptotic properties

In this Section, we present the mean and variance asymptotical calculations of $\hat{\rho}(\mathbf{x})$ in the case of locally fitted generalized linear model.

Ruppert & Wand (1994) have derived the asymptotic properties of bias and variance terms for general multivariate kernel weights using weighted least squares matrix theory. Fan et al. (1995) have investigated the extension of the nonparametric regression technique of local polynomial fitting with a kernel weight to generalized linear models. Staniswalis (1989a) carried out a similar generalization of the Nadaraya-Watson kernel estimator (Nadaraya 1964; Watson 1964), which is equivalent to local fitting with a kernel weight. Signorini & Jones (2004) have presented a rather thorough investigation of the use of kernel based nonparametric estimators of the binary regression function in the case of a single covariate.

Fan (1992a, 1993) and Ruppert & Wand (1994) showed that in the ordinary regression context, local polynomial kernel regression has many attractive mathematical properties. This is particularly true when the polynomial is of odd degree, because the asymptotic bias close to the boundary of the support of the covariates can be shown to be of the same order of magnitude as that of the interior. This is not true

for the Nadaraya-Watson kernel estimator, because it corresponds to the zero degree fitting. In addition, the asymptotic bias of odd degree polynomial fits at a point $\mathbf{x}$ depends on $\mathbf{x}$ only through a higher order derivative of the regression function itself, which allows for simple interpretation and expressions for the asymptotically optimal bandwidth. They have showed that these properties carry over to generalized linear models. See Fan et al. (1995) for more details.

In order to obtain the asymptotic properties of bias and variance of $\hat{\rho}$, we need to make the following assumptions.

(i) The kernel K is a spherically symmetric continuous density function with bounded support such that $\int \mathbf{u}\mathbf{u}^T K(\mathbf{u})\mathbf{d}\mathbf{u} = \mu_2(K)I$, where $\mu_2(K) \neq 0$ is scalar and I is the identity matrix of degree 2. As K is spherically symmetric kernel, all odd moments of K vanish, that is, $\int u_1^{l_1} u_2^{l_2} K(\mathbf{u})d\mathbf{u} = 0$ for all non-negative integers l_1 and l_2 such that their sum is odd.

(ii) The bandwidth $h \to 0$ and $nh^2 \to \infty$ as $n \to \infty$ ($n = n_1 + n_2$), and that $\dfrac{n_1}{n_2}$ is fixed.

(iii) The point $\mathbf{x}$ is interior to $\mathcal{R}$.

(iv) $f(\mathbf{x}) > \epsilon$ and $g(\mathbf{x}) > \epsilon$ for all $\mathbf{x} \in \mathcal{R}$.

We consider here only the linear case of the polynomial regression. That is,

$$\log\left[\frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})}\right] = \hat{\beta}_0 + \hat{\beta}_1(z_1 - x_1) + \hat{\beta}_2(z_2 - x_2) \tag{4.4.1}$$

Theorem 4.4.1. *Assume that the conditions (i)-(iv) are satisfied. Then*

$$E\{\hat{\rho}_{LL}(\boldsymbol{x}; h)\} = \rho(\boldsymbol{x}) + \frac{1}{2}h^2\mu_2(K)\nabla^2\rho(\boldsymbol{x}) + o(h^2). \tag{4.4.2}$$

and

$$Var\{\hat{\rho}_{LL}(\boldsymbol{x}; h)\} = h^{-2}R(K) \quad \left[\frac{1}{n_1 f(\boldsymbol{x})} + \frac{1}{n_2 g(\boldsymbol{x})}\right] + o\left(\frac{1}{n_1 h^2} + \frac{1}{n_2 h^2}\right). \tag{4.4.3}$$

These results can be derived from the theorems in Fan et al. (1995). A direct proof of the above theorem is given in the Appendix B.

These expressions of bias and variance of local linear estimator have a simple interpretation. The leading bias term (4.4.2) depends $\mathbf{x}$ only through $\nabla^2\rho(\mathbf{x})$ which reflects the error of the linear approximation. If ρ is close to being linear at $\mathbf{x}$ then $\nabla^2\rho(\mathbf{x})$ is relatively small, which is consistent with the concept of local fits having less bias in this case. On the other hand, if ρ has a high amount of curvature at x then $\nabla^2\rho(\mathbf{x})$ is higher and local linear fits tend to produce more biased estimates. The fact that the bias depends on h^2 also reflects the idea that bias is increased with more smoothing. Indeed, the expression 4.4.2 is a direct analogue of the bias approximation for kernel density estimation given by 2.4.4 in Chapter 2.

Comparing the leading local linear (LL) bias term of $\hat{\rho}$ in equation (4.4.2) to (2.4.5), bias estimator for the density ratio method, we see that the local linear method is unbiased at the points where ρ is linear, whereas the density ratio method is generally biased in that situation. This helps to explain why we might expect the local linear method to be preferable when the trend is large scale and relatively slow varying.

We note that the asymptotic variance of LL estimator (4.4.3) is the same as for the density ratio estimator (2.4.6). As expected, it produces high variability at the locations where the population density is extremely low.

4.5 Methods of bandwidth selection

The bandwidth of course has a crucial effect on the quality of the kernel smoothing estimators. As mentioned in Chapter 3, it can be chosen either subjectively by users or objectively by the data. One can opt for a fixed bandwidth also referred to as a estimation-point adaptive bandwidth or a sample point adaptive bandwidth.See Jones (1990). The theoretical selection of a variable bandwidth was discussed in Fan & Gijbels (1992) in the context of standard regression in detail. Here in this chapter, we do not elaborate on this but focus on the choice of a fixed bandwidth.

In principle both the cross-validation methods described in Chapter 3 can be used with the local linear estimator $\hat{\rho}_{LL}$. However, through a simulation study (in the section 4.6.2), we found that the results are disappointing. This motivates us to develop a new plug-in bandwidth selection method as follows.

4.5.1 Plug-in bandwidth selector

The performance of the relative risk estimator (in log scale) require the specification of appropriate error criteria for measuring the error when estimating the density at a single point as well as the error when estimating the density over the region. So as usual, we consider an error criterion that globally measures the distance between the functions $\hat{\rho}(\mathbf{x})$ and $\rho(\mathbf{x})$. Recall that the mean and variance in the asymptotic properties of $\hat{\rho}$ for LL method are

$$\mathbb{E}[\hat{\rho}_{LL}(\mathbf{x})] = \rho(\mathbf{x}) + \frac{h^2}{2}\mu_2(K)\nabla^2\rho(\mathbf{x}) + o(h^2) \tag{4.5.1}$$

and

$$\mathrm{Var}[\hat{\rho}_{LL}(\mathbf{x})] = \frac{R(K)}{h^2}\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right] + o(n_1^{-1}h^{-2} + n_2^{-1}h^{-2}). \tag{4.5.2}$$

Therefore,

$$\begin{aligned} MISE(h) &\approx \int_{\mathcal{R}} \left[\frac{h^2}{2}\mu_2(K)\nabla^2\rho(\mathbf{x})\right]^2 d\mathbf{x} + \int_{\mathcal{R}} \frac{R(K)}{h^2}\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right]\mathbf{dx}. \\ &\approx \frac{h^4}{4}\mu_2(K)^2 \int_{\mathcal{R}} \left[\nabla^2\rho(\mathbf{x})\right]^2 \mathbf{dx} + \frac{R(K)}{h^2}\int_{\mathcal{R}}\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right]\mathbf{dx}. \end{aligned} \tag{4.5.3}$$

The integrated variance (second term of 4.5.3) may become huge if f or g are very small in some parts of the study region. In such cases, minimization of MISE(h) will produce a very large bandwidth in order to stabilize the estimate where the data are sparse, at the expense of oversmoothing elsewhere. An alternative is to use a population weighted version of MISE,

$$\mathrm{WMISE}(h) = \int_{\mathcal{R}} [\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})]^2 g(\mathbf{x}) d\mathbf{x}.$$

See Hazelton (2008) for more details. This performance criterion will obviously be similar to MISE when the controls are distributed in a fairly uniform manner over $\mathcal{R}$. When the control distribution is markedly heterogeneous then WMISE reduces (in comparison to MISE) the weight given to errors in areas where the (control) data are sparse.

A standard asymptotic expansion of WMISE(h) leads to the approximation

$$WMISE(h) \approx \frac{h^4}{4}\mu_2(K)^2 \int_{\mathcal{R}} [\nabla^2 \rho(\mathbf{x})]^2 g(\mathbf{x}) d\mathbf{x} + \frac{R(K)}{h^2} \int_{\mathcal{R}} \left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right] g(\mathbf{x}) d\mathbf{x}. \tag{4.5.4}$$

A problem in developing a plug-in bandwidth selector based on this expression is that it is difficult to obtain stable estimates of the integrated variance (IV), the second term of the equation (4.5.4) because of the need for a pilot estimate of the reciprocal of f. We therefore make a further assumption that $f(\mathbf{x}) \approx g(\mathbf{x})$ over most of $\mathcal{R}$, that is, the relative risk is close to one for the most part, then the IV becomes,

$$\begin{aligned} IV &\approx h^{-2} R(K) \int_{\mathcal{R}} \left[\frac{1}{n_1} + \frac{1}{n_2}\right] \frac{1}{g(\mathbf{x})} \quad g(\mathbf{x})\mathbf{dx} \\ &\approx h^{-2} R(K) \left[\frac{1}{n_1} + \frac{1}{n_2}\right] \mid \mathcal{R} \mid \end{aligned}$$

where $\mid \mathcal{R} \mid$ is the area of the region.

We can estimate the integrated squared bias, ISB (first term of equation 4.5.4) by replacing ρ with a pilot estimate thereof, and changing integration with respect to g by averaging over the observed control data. We then get the estimate

$$\widehat{ISB} = \frac{h^4}{4}\mu_2(K)^2 \frac{1}{n_2} \sum_{i=1}^{n} (1 - y_i)[\nabla^2 \bar{\rho}(\mathbf{x}_i)]^2 \tag{4.5.5}$$

where $\bar{\rho}$ is a pilot estimate of ρ. In practice, we obtain this using the density ratio estimator with bandwidth $5h_{OS}$, where h_{OS} is the oversmoothing bandwidth of Terrell (1990). This would be a very large bandwidth for estimating a density *per se*, but we are interested here in estimating a functional of derivatives of densities, which require much greater amounts of smoothing (e.g. Wand & Jones, 1995). $\nabla^2\bar{\rho}$ is computed using numerical differentiation. Our precise choice of pilot bandwidth is ad hoc, but it performs well in numerical experiments (and uses in the simulation study, to be presented in the next Section).

We apply the approximations for IV and ISB, then we get

$$WMISE(h) \approx \frac{h^4}{4}\mu_2(K)^2\frac{1}{n_2}\sum_{i=1}^{n}(1-y_i)[\nabla^2\bar{\rho}(\mathbf{x}_i)]^2 + h^{-2}R(K)\left[\frac{1}{n_1}+\frac{1}{n_2}\right]|\mathcal{R}|$$

By differentiating this with respect to h and setting the derivative equal to zero,

$$0 = h^3\mu_2(K)^2\frac{1}{n_2}\sum_{i=1}^{n}(1-y_i)(\nabla^2\rho(\mathbf{x}_i))^2 - \frac{2R(K)|\mathcal{R}|}{h^3}[(n_1^{-1}+n_2^{-1})]$$

Hence our plug-in bandwidth selector ($h_{LL.PI}$) for the local linear method is

$$\hat{h}_{LL.PI} = \left[\frac{2|\mathcal{R}|R(K)(n_1^{-1}+n_2^{-1})}{\mu_2(K)^2\frac{1}{n_2}\sum_{i=1}^{n}(1-y_i)\left[\nabla^2\bar{\rho}(\mathbf{x}_i)\right]^2}\right]^{1/6}.$$

Note that the assumption $f(\mathbf{x}) \approx g(\mathbf{x}), \mathbf{x} \in \mathcal{R}$ violates when there is a large scale trend. We can estimate ISB as appeared in equation (4.5.5), but the estimate for IV should be changed. Then IV from equation (4.5.4) becomes,

$$IV \approx h^{-2}R(K)\int_{\mathcal{R}}\left[\frac{1}{n_1 f(\mathbf{x})}+\frac{1}{n_2 g(\mathbf{x})}\right] \quad g(\mathbf{x})\mathbf{dx}.$$

When estimating this, it is hard to get the stable estimates for IV since it contains the reciprocal of f. Therefore, we further assume that $f(\mathbf{x}) > 0$, for every $\mathbf{x} \in \mathcal{R}$. We can estimate this by replacing f and g with pilot estimates $\bar{f}$ and $\bar{g}$, then changing the integration with respect to g by averaging over the observed control data.

4.6 Simulation study to compare local linear against density ratio estimator

In order to explore the practical performance of density ratio estimator $\hat{\rho}$ with local linear estimator $\breve{\rho}_{LL}$, we present the results of two simulation studies. In the first study, the optimal smoothing is used to compute both estimators while a data-driven bandwidth is used in the latter study.

4.6.1 Simulation results: with optimal smoothing

We consider four synthetic problems (a)-(d), containing different case (f) and control (g) densities over the square region $\mathcal{R}$=[0,10] $\times$ [0,10]. Table 4.2 provides definitions of case and control densities, while contour plots of the densities and corresponding log-relative risk functions are depicted in Figure 4.1. Problems (a) and (c) are extreme settings in terms of the pattern of risk, with the former involving a global linear trend in ρ over $\mathcal{R}$, and the latter containing two local disease hot spots. The trend in problem (d) is also rather large scale in nature, while in Problem (b) the function ρ has a more complex local structure.

Table 4.2: Case and control densities for four synthetic problems. The function ϕ_σ is a bivariate normal density with zero mean vector and covariance matrix $\sigma^2 I$, where I is the 2×2 identity matrix. In addition, $\phi_\sigma^{\mathcal{R}}$ denotes ϕ_σ truncated to $\mathcal{R}$. The location parameters are $\mu_1 = [5,5]^T$, $\mu_2 = [8,6]^T$, $\mu_3 = [3,3]^T$ and $\mu_4 = [5,4]^T$.

Problem (a)	$f(z) = \frac{1}{600}(z_1 + 1)$ $g(z) = 0.01$
Problem (b)	$f(z) = 0.001 + 0.7\phi_2^R(z-\mu_1) + 0.1\phi_2^R(z-\mu_2) + 0.1\phi_1^R(z-\mu_3)$ $g(z) = 0.001 + 0.4\phi_2^R(z-\mu_1) + 0.5\phi_2^R(z-\mu_2)$
Problem (c)	$f(z) = 0.001 + 0.7\phi_3^R(z-\mu_1) + 0.1\phi_{0.5}^R(z-\mu_2) + 0.1\phi_{0.5}^R(z-\mu_3)$ $g(z) = 0.001 + 0.9\phi_3^R(z-\mu_1)$
Problem (d)	$f(z) = 0.001 + 0.6\phi_3^R(z-\mu_1) + 0.3\phi_2^R(z-\mu_4)$ $g(z) = 0.01$

For each problem, we considered sample sizes $n_1 = n_2 = 100$ and $n_1 = n_2 = 500$. We generated 400 case-control data sets for each combination of problem and sample size, and then estimated the log-relative risk function using both density ratio and local linear methods.

The estimation methods were implemented using optimal bandwidths in all cases. These were computed through a pilot study in which we found the bandwidth for each method and sample size which minimizes the mean integrated squared error of the log-relative risk estimates over the pilot data sets. They are tabulated in Table 4.3. As a result, our methods in the simulation study are operating with unrealistically good bandwidths, but we have controlled for differences in the performance

Table 4.3: Optimal bandwidths for the Problems for density ratio (DR) and local linear (LL) estimators.

	Data	DR	LL
Problem (a)	$n = 100$	1.97	6.02
	$n = 500$	1.28	2.96
Problem (b)	$n = 100$	2.34	6.58
	$n = 500$	1.52	2.11
Problem (c)	$n = 100$	6.25	6.40
	$n = 500$	1.52	6.34
Problem (d)	$n = 100$	1.42	2.65
	$n = 500$	0.89	2.07

Table 4.4: MISE estimates in the simulation study to compare DR over LL methods.

	Data	DR	LL
Problem (a)	$n = 100$	52.74	36.56
	$n = 500$	25.46	15.18
Problem (b)	$n = 100$	51.28	117.68
	$n = 500$	32.76	53.06
Problem (c)	$n = 100$	43.71	70.07
	$n = 500$	41.85	48.62
Problem (d)	$n = 100$	141.25	88.34
	$n = 500$	82.46	33.82

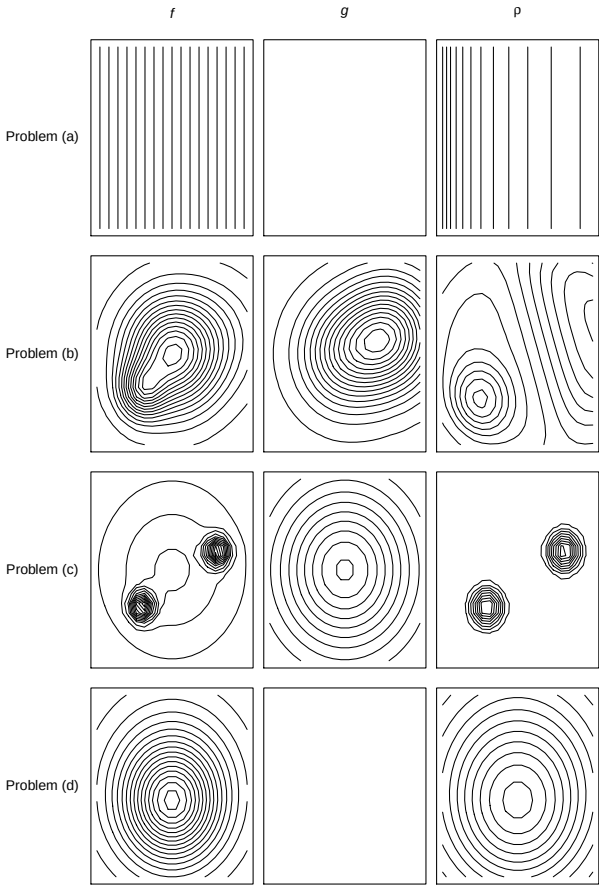

Figure 4.1: Contour plots of case density f, control density g, and log relative risk function ρ for the four synthetic problems.

of the bandwidth selectors for both methods. We computed the integrated squared error (ISE) for each estimate of ρ. MISE estimates are given in Table (4.4) and the results are displayed (on the log-scale) as boxplots in Figure 4.2, where DR and LL distinguish density ratio from local linear results, and where 1 and 5 distinguish $n_1 = n_2 = 100$ from $n_1 = n_2 = 500$ cases.

Overall, the results match our expectations well. When the trend in the log-relative

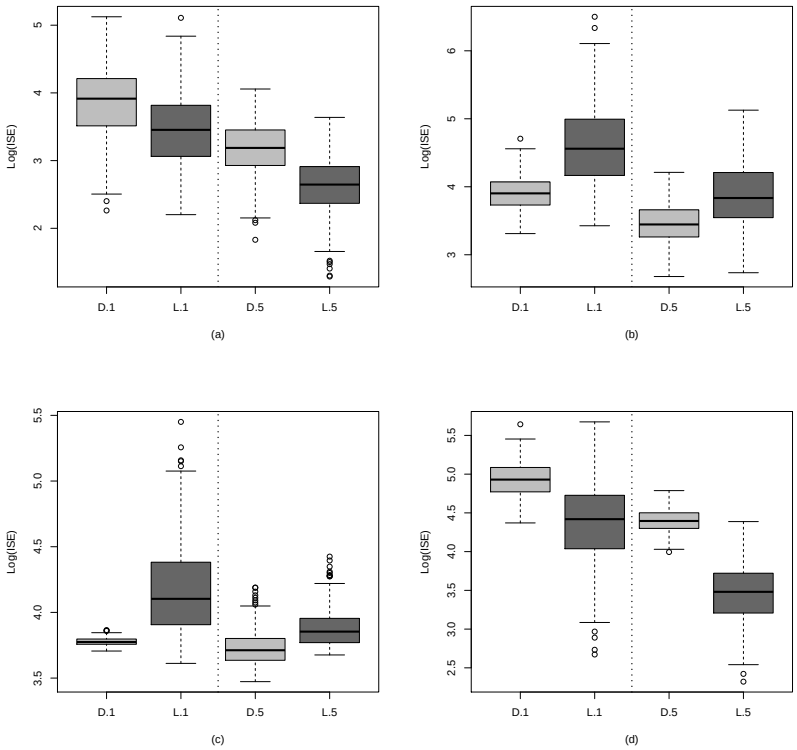

Figure 4.2: Boxplots of log(ISE) for DR and LL estimates of ρ. The suffix 1 indicates sample sizes $n_1 = n_2 = 100$ and similarly for 500.

risk varies relative slowly over $\mathcal{R}$, as in Problems (a) and (d), the local linear estimator is to be preferred. In some cases the advantage of the local linear estimator is substantial, it improves over the density ratio estimator by about three times the median ISE for the larger sample size in Problem (d), for example. In Problems (b) and (c), where the trend in ρ has a more complex local structure, the density ratio method performs better than the local linear. This is most likely due to the presence of circular risk hotspots. Also LL produces an over smoothing estimator while DR an under smoothing estimator during this estimation.

Identifier	Description	Formula
1	Mild circular clustering:	$r(\mathbf{z}) \propto 1 + e^{-\theta} \parallel \mathbf{z} - [5,5]^{\mathsf{T}} \parallel$
2	Strong circular clustering:	$r(\mathbf{z}) \propto 1 + 4 \times e^{-\theta} \parallel \mathbf{z} - [5,5]^{\mathsf{T}} \parallel$
3	Mild linear clustering:	$r(\mathbf{z}) \propto 1 + 1 \times e^{-\theta} \parallel z_2 - 5 \parallel$
4	Strong linear clustering:	$r(\mathbf{z}) \propto 1 + 4 \times e^{-\theta} \parallel z_2 - 5 \parallel$

Table 4.5: Relative risk functions for test problems.

4.6.2 Simulation results: with data-driven bandwidths

In this study, we omit LSCV bandwidth selector in the estimation of ρ_{LL} since it is a worse performer than that of LCV in preliminary study. Therefore we use LCV and plug-in (PI) bandwidth selector in the LL estimation of relative risk function. LSCV is used to compute density ratio estimator.

In this study, we contrast the finite sample performance of the local linear estimator $\breve{\rho}_{LL}$ with that of the density ratio estimator $\hat{\rho}$. We consider sixteen synthetic problems, generated by a 2^4 factorial design. The first two factors are the type of clustering (circular or linear) and strength of clustering (mild or strong). The corresponding forms of the relative risk functions are displayed in Table 4.5, where the region $\mathcal{R}$ is defined to be the square $[0, 10] \times [0, 10]$. The next factor that we consider is the scale of the cluster, which is defined by the parameter θ in the expressions in the Table. Specifically, we examine the cases $\theta = 1$ (short range) and $\theta = 0.5$ (wide range). Greyscale plots of the relative risk functions from Table 4.5 when $\theta = 1$ are presented

in Figure 4.3. Our final factor is the form of the control density, which is either uniform on $\mathcal{R}$ or bivariate normal density with mean $\mu = (5, 5)^{\mathsf{T}}$ and covariance matrix $\Sigma = 2I$ (truncated at the boundaries of $\mathcal{R}$). Here we consider the control densities in a single mode, the results provide guidance to cases where multiple modes are very well separated as we might expect to see with city and towns with real applications. Risk surface 1 involves a mild circular clustering over the region $\mathcal{R}$ whilst a strong circular clustering is displayed in risk surface 2. There is a linear clustering in risk surfaces 3 and 4 respectively varying from low to higher.

For each problem, we considered sample sizes $n_1 = n_2 = 100$ and $n_1 = n_2 = 400$. Since the computation of LL estimator is computationally expensive, we restrict the number of iterations to 100. Then we estimate the log-relative risk function using density ratio estimators implemented with the bandwidth $\hat{h}_{DR.CV}$, and local linear estimators implemented with bandwidths $\hat{h}_{LL.LCV}$ and $\hat{h}_{LL.PI}$. We computed the integrated squared error (ISE) for each estimate of ρ.

The results from 100 case-control data sets for each problem scenario are summarized (on the log-scale) in the form of boxplots in Figures 4.4, 4.5, 4.6 and 4.7. In these plots DR and LL distinguish the density ratio from local linear results and CV, LCV and PI indicate least squares cross validation, likelihood cross validation and plug-in bandwidth estimators. The average of ISE of log relative risk is displayed in Table 4.6.

Perhaps the most striking feature of the results is the poor performance of the local-linear method implemented using the likelihood-cross validation bandwidth. The

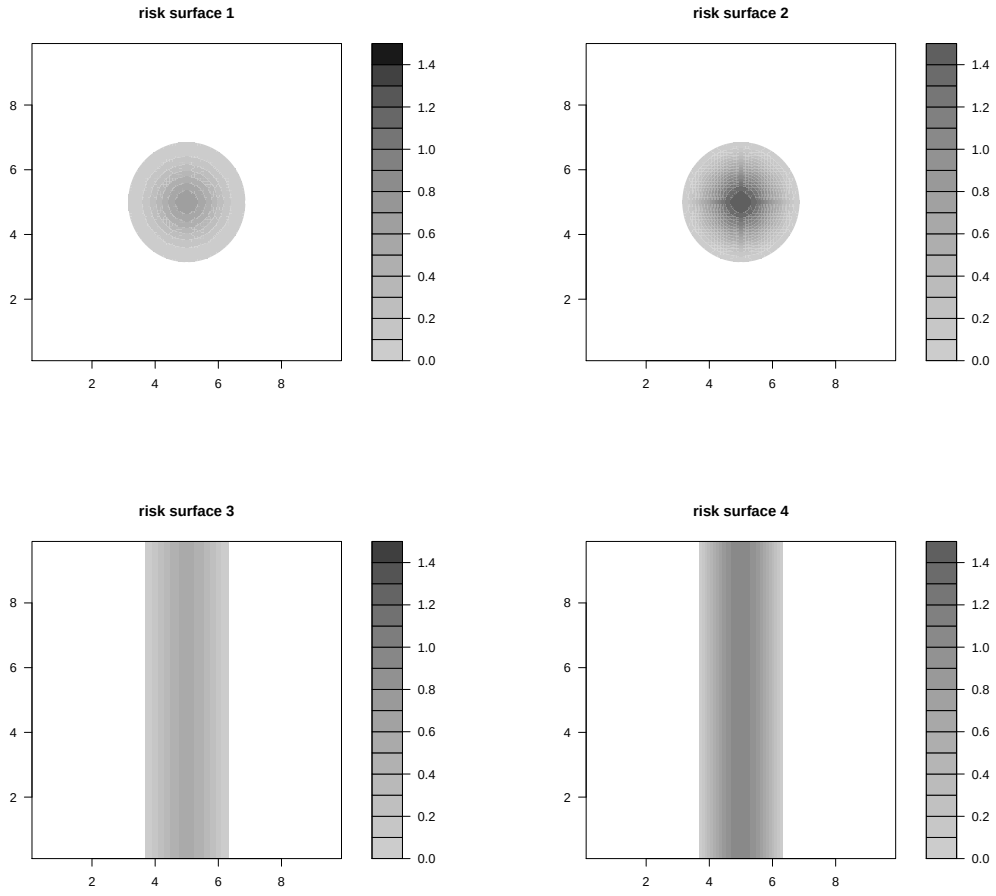

Figure 4.3: Filled contour plots for the test relative risk functions (on the log scale).

local-linear method fares much better with our plug-in bandwidth, and produces the smallest estimation errors in some of the scenarios with linear risk sources. Nonetheless, the density ratio estimators perform better in more than half the scenarios, including all cases with circular risk clusters and (perhaps surprisingly) some with linear risk clusters. The reason for the poor performance of LL estimator may be due to the optimal bandwidth. This bandwidth is inversely proportional to a second

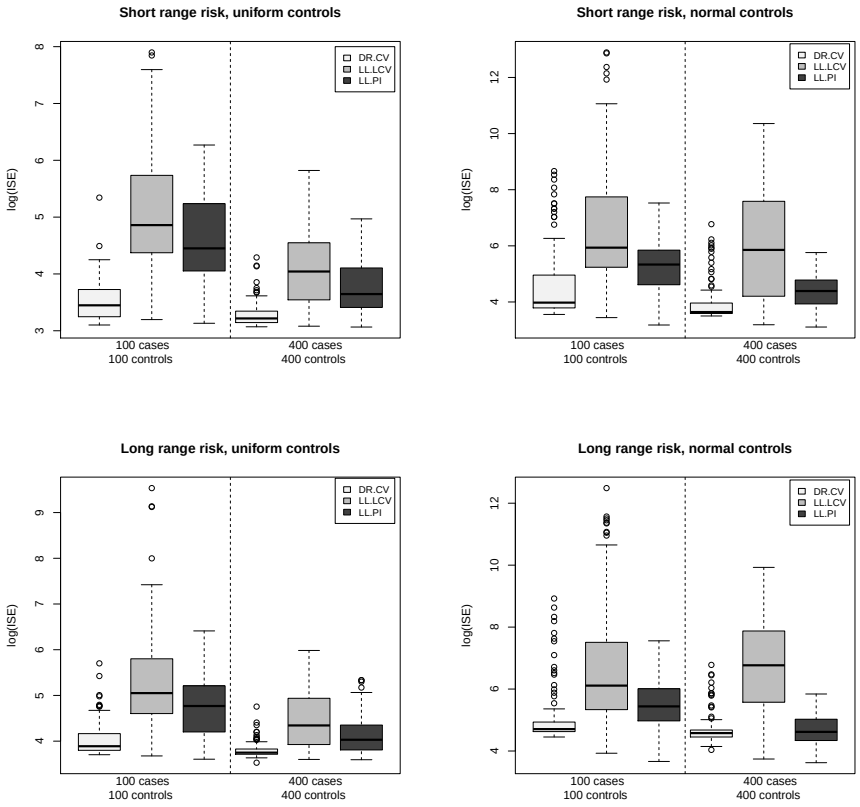

Figure 4.4: Boxplots of log(ISE) for estimates of the log-relative risk for test problem 1 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text.

derivative term, which seems complex in the estimation. So we compute it numerically.

However, it is important to note that with all the cross validation bandwidth selectors, the chosen value of h occurred at a boundary of the search interval $[h_0, 4h_0]$ in over 25% of the simulated data sets in some scenarios. This indicates a somewhat worrying degree of variability in the cross-validation bandwidth selectors.

Table 4.6: MISE estimates in the simulation study to compare DR over LL methods. Medians are displayed inside brackets.

Problem:	Data	DR—CV	LL—LCV	LL—PI
1	100	37.2 (31.4)	286.7 (128.9)	133.7 (85.6)
	400	27.6 (24.9)	82.2 (57.0)	48.8 (38.3)
2	100	179.9 (171.2)	1549.6 (356.9)	259.9 (235.5)
	400	151.3 (156.6)	245.0 (204.6)	165.1 (159.0)
3	100	140.8 (132.3)	1785.8 (269.1)	219.6 (183.2)
	400	115.4 (121.9)	167.7 (139.0)	134.4 (128.5)
4	100	697.9 (687.3)	1.4E+72 (1488.1)	805.3 (768.9)
	400	495.8 (499.7)	497.9 (390.9)	557.6 (558.3)
5	100	398.5 (53.3)	16325.4 (378.3)	282.6 (207.1)
	400	80.4 (38.1)	1694.0 (349.7)	93.8 (80.7)
6	100	1660.6 (426.4)	66603.9 (12764.4)	602.2 (496.1)
	400	367.7 (296.9)	12494.8 (8067.3)	324.4 (276.2)
7	100	478.7 (188.3)	18099.6 (1050.2)	427.5 (326.4)
	400	198.6 (171.6)	3003.6 (1227.8)	210.8 (196.2)
8	100	2647.1 (1384.9)	132497.9 (70218.9)	1089.3 (1055.3)
	400	849.7 (795.7)	23987.2 (12056.3)	548.2 (528.2)
9	100	62.3 (48.9)	640.7 (156.1)	146.3 (117.7)
	400	45.8 (42.4)	113.5 (77.0)	68.1 (56.3)
10	100	302.8 (300.5)	2614.1 (587.3)	365.9 (329.0)
	400	204.3 (200.5)	275.9 (246.5)	236.6 (234.4)
11	100	158.1 (154.6)	664.4 (269.5)	246.6 (209.4)
	400	123.3 (121.6)	190.4 (153.4)	147.7 (145.5)
12	100	731.3 (679.5)	6501.5 (2146.8)	866.7 (843.7)
	400	446.0 (436.1)	487.3 (371.1)	533.8 (522.5)
13	100	425.1 (110.7)	11196.0 (450.9)	313.1 (230.5)
	400	134.9 (97.3)	2251.9 (870.3)	121.1 (101.2)
14	100	2609.9 (794.6)	93685.5 (42137.1)	718.5 (612.5)
	400	562.2 (432.4)	18388.2 (10624.9)	445.1 (355.2)
15	100	665.2 (251.9)	22404.4 (714.1)	432.7 (375.9)
	400	241.6 (225.0)	3607.1 (1152.3)	200.8 (193.4)
16	100	3098.9 (1661.3)	83273.3 (25256.6)	1189.6 (1219.7)
	400	813.4 (773.9)	18122.8 (12432.5)	494.4 (467.9)

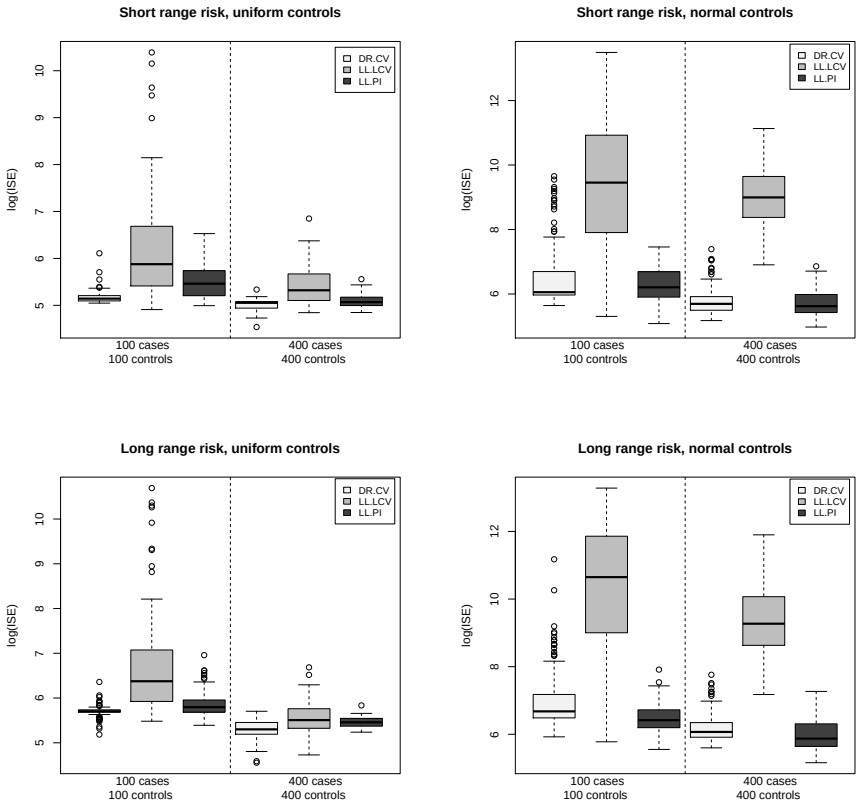

Figure 4.5: Boxplots of log(ISE) for estimates of the log-relative risk for test problem 2 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text.

4.7 Tolerance contours of local linear estimators

The benefits of kernel smoothing based relative risk functions in exploratory analysis and data visualization is enhanced if one is able to highlight the areas of significantly elevated risk. In this section, we present tolerance contours to distinguish significant features of the relative risk function for the local linear method. As before, we use p-values derived from z-tests for constructing tolerance contours for the local linear

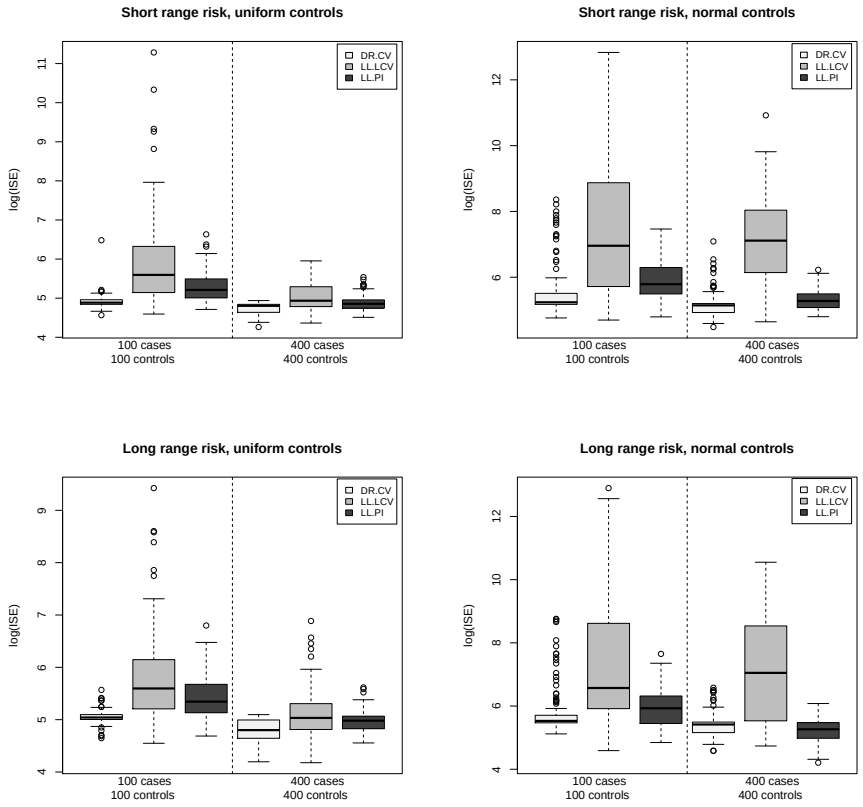

Figure 4.6: Boxplots of log(ISE) for estimates of the log-relative risk for test problem 3 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text.

estimator be obtained by using the statistic

$$Z(\mathbf{x}) \;=\; \frac{\hat{\rho}(\mathbf{x})}{SE\{\hat{\rho}_{LL}(\mathbf{x})\}},$$

where the standard error of $\hat{\rho}_{LL}(\mathbf{x})$ is

$$SE\{\hat{\rho}_{LL}(\mathbf{x})\} = \sqrt{\frac{1}{h^2} R(K) \frac{1}{\hat{g}_p(\mathbf{x})} \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}.$$

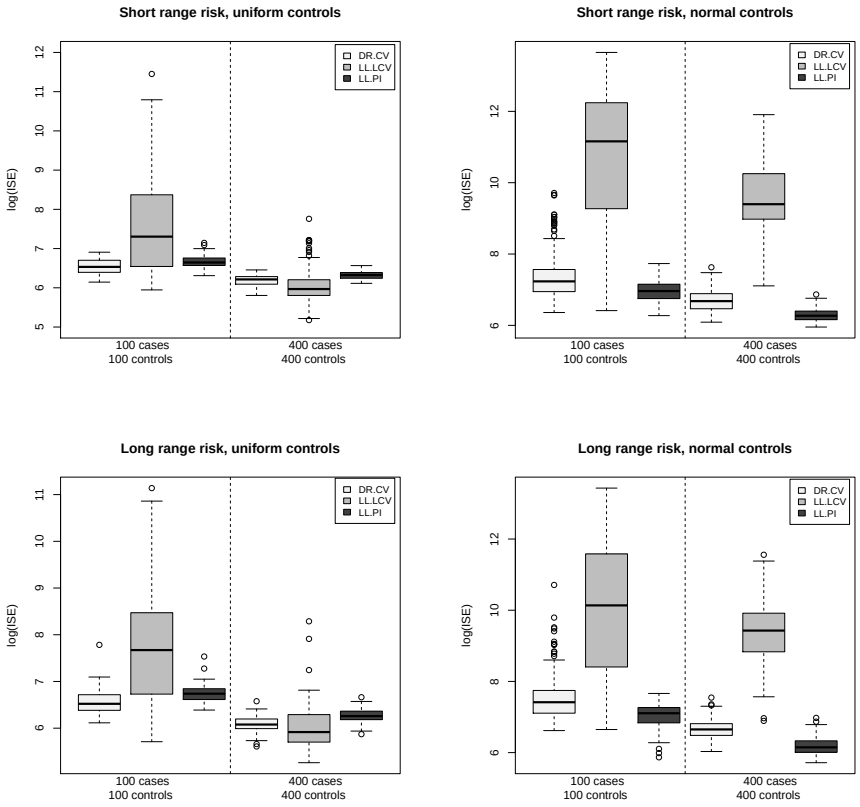

Figure 4.7: Boxplots of log(ISE) for estimates of the log-relative risk for test problem 4 from Table 1. Short and long ranges indicate values $\theta = 1$ and $\theta = 0.5$ respectively; the control density is specified as uniform or normal as described in the text.

From equation (4.5.2) with the additional assumption (from H_0) that $f = g$. The density estimate $\hat{g}_p(\mathbf{x})$ is a pooled estimator, constructed from a combined data set comprising both cases and controls. We note that if $\mathbf{x}$ is close to the boundary of $\mathcal{R}$ then it is necessary to edge correct the density estimator $\hat{g}_p(\mathbf{x})$, and to replace $R(K)$ by $\int_{\mathcal{R}} \{K_h(\mathbf{x})\}^2 \, d\mathbf{x}$. The test statistic Z has an asymptotic standard normal distribution under H_0 (Fan et. al, 1995), and so we may hence compute the requisite field of p-values over $\mathcal{R}$.

4.8 Real applications

In this Section, we illustrate the estimation of spatial relative risk in three particular applications using the local linear and density ratio methods.

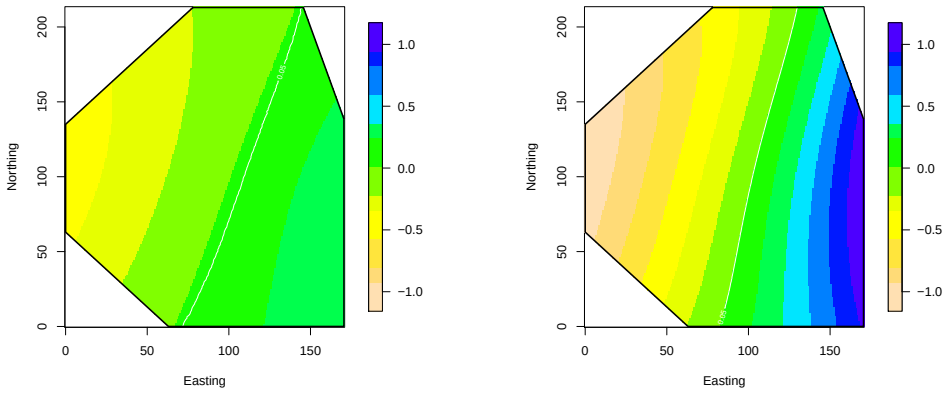

Figure 4.8: Estimates of the log-relative risk of disease from the Myrtle Beech data. The left-hand panel shows the estimate using the density ratio method with least-squares cross-validation bandwidth $h = 64$, and the right-hand one the local linear estimator using our plug-in bandwidth $h = 197.3$. The dashed lines indicate 95% tolerance contours for areas of elevated risk.

4.8.1 Myrtle tree data

The first application that we analyze is related to the disease Myrtle Wilt for trees growing in Tasmania, Australia, as discussed in Chapter 2. There are a number of spatial models that might be applied to data such as these. Our aim is to estimate the relative risk function (along with associated tolerance contours) as a tool for exploratory data analysis.

We computed log-relative risk estimates on data using both density ratio and local linear methods. Bandwidths 64.0 and 197.3 respectively were obtained using the least squares cross-validation and plug-in methods respectively. The resulting estimates are displayed as contour plots in Figure 4.8, along with 95% tolerance contours. The local linear method is arguably preferable in this case. Certainly it is noticeable that the density ratio method indicates a rather shallow slope in the risk surface from left to right (corresponding to a difference of less that one unit on the log-relative risk scale from left hand to right edges), whereas the trend seems rather stronger than this when one eyeballs the data. An explanation is that a large bandwidth is required for the density ratio method in order to produce a sufficiently smooth estimate, but the resulting bias flattens the gradient of $\hat{\rho}$.

4.8.2 Chorley-Ribble cancer data

Our second example is concerned with cancer data in Chorley Ribble as discussed earlier.

As in the previous example, we estimated the log-relative risk functions using both density ratio and local linear methods. Bandwidths 0.78 and 16.66 respectively were obtained using the least squares cross-validation for DR method and plug-in for LL method respectively. The resulting estimates are displayed in Figure 4.9. The two methods provide rather different interpretations of the data. The density ratio approach indicates modest fluctuations in the relative risk over the region, none of which

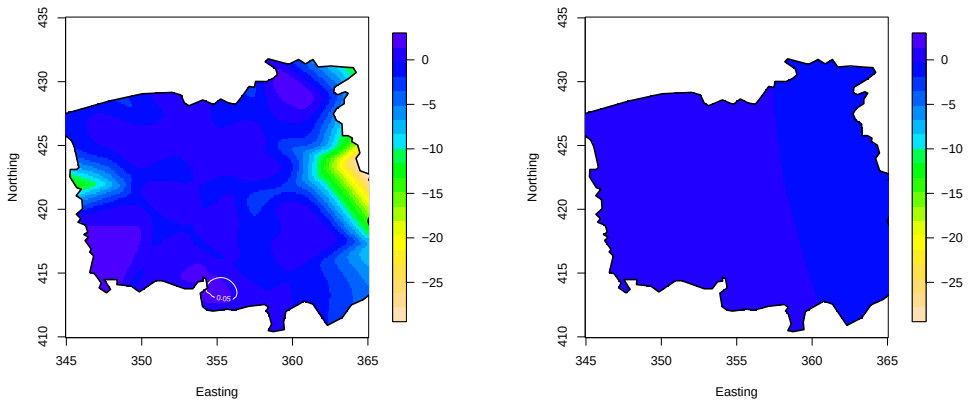

Figure 4.9: Estimates of the log-relative risk of larynx cancer in the Chorley-Ribble region of Lancashire, England. The left-hand panel shows the estimate using the density ratio method with least-squares cross-validation bandwidth $h = 0.78$, and the right-hand one the local linear estimator using our plug-in bandwidth $h = 16.66$. The dashed lines indicate 95% tolerance contours for areas of elevated risk.

are significant (in terms of tolerance limits) except for the areas close to the aforementioned incinerator. The local linear method, on the other hand, suggests a very flat relative risk profile over the study region. It could be argued that a real effect of the incinerator has been missed in this case.

4.8.3 Foot and mouth disease (FMD) data

The third data set we use is from foot-and-mouth disease (FMD) data from the 1967 outbreak in Cheshire, UK. FMD is a highly infectious viral disease of farm livestock. The virus can be spread directly between animals over short distances in contaminated airborne droplets, and indirectly over longer distances, for example via the movement of contaminated material.

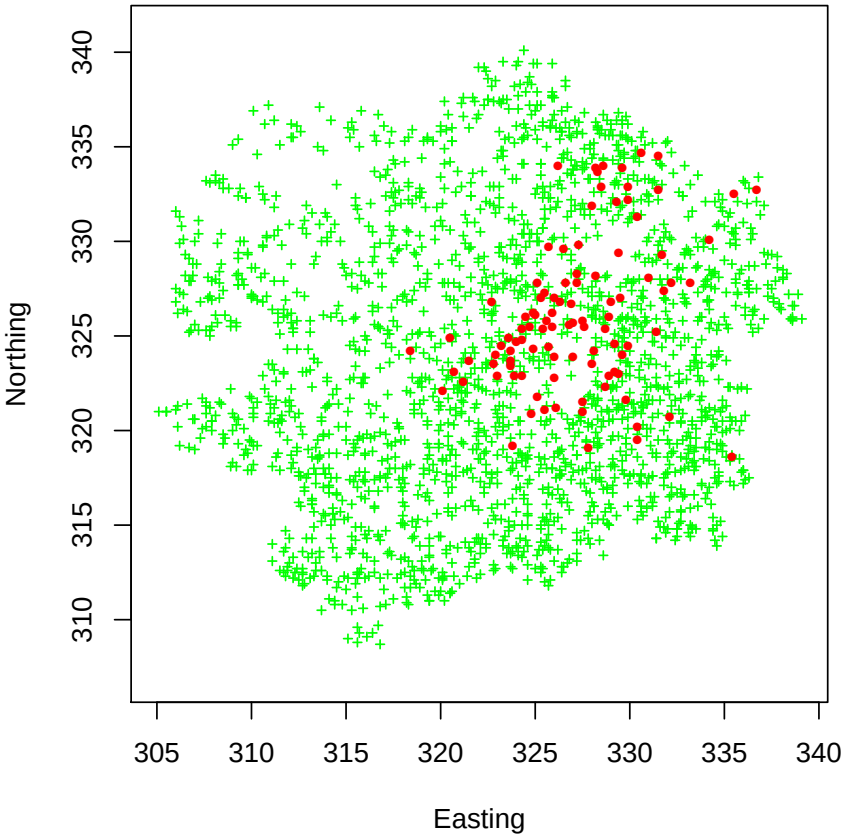

Figure 4.10: Spatial distribution of FMD data. 100 cases (•), 2129 controls (+).

The spatial distribution of reported cases and controls of FMD can be seen in Figure 4.10. The cases comprise 100 individual with FMD confirmed from daily inspections. There are 2129 control farms thought to be at high risk of acquiring the disease. These data have been collected over a period of 16 days.

As in both previous data sets, we investigate the spatial distribution of this data set, compute the relative risk surface and hence highlight the farms which correspond to significantly high or low risk using the 95% tolerance contours using the

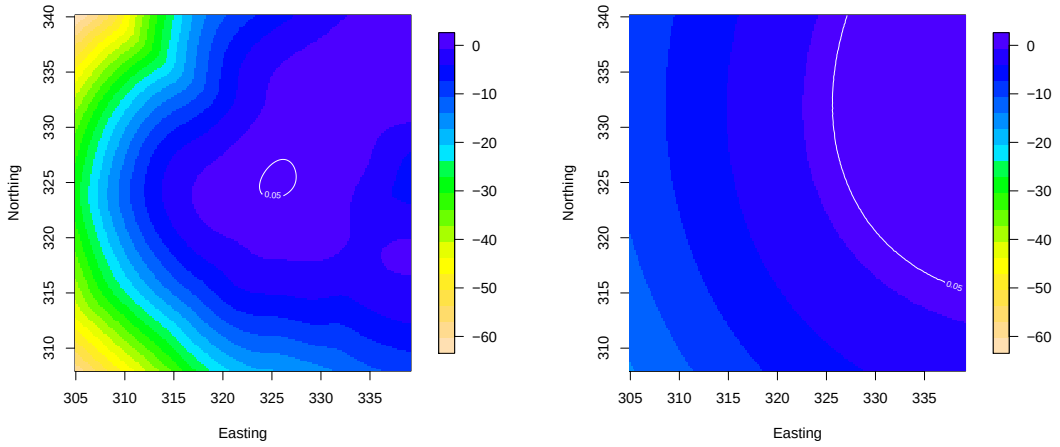

Figure 4.11: These Figures show the estimates of log-relative risk of disease from FMD data. The left panel is used DR method with LSCV bandwidth $h = 1.74$, while the right panel shows the estimate using LL method with PI bandwidth $h = 13.65$. The dashed lines indicate 95% tolerance contours for areas of elevated risk.

DR and LL methods with the help of LSCV and PI bandwidth selectors respectively. Bandwidths 1.74 and 13.65 were respectively obtained using LSCV and PI bandwidth selectors for DR and LL methods respectively. The resulting DR (left panel) and LL estimates (right panel) are displayed in Figure 4.11. Tolerance contours for both methods indicate the regions at central to northeastern region exhibit the most risk. The DR approach produces a risk surface showing significance in the middle of the region while LL suggests a flatter risk surface compared to the former estimator over the study region. The local linear method is arguably oversmoothing in this case.

4.9 Conclusion

In this chapter, we have compared and contrasted the density ratio and local linear estimators of the log-relative risk function in geographical epidemiology. In previous studies (e.g. Clark & Lawson, 2004) the local-linear approach has appeared highly competitive for estimating the conditional probability function. However, our results overall suggest that the density ratio method is generally better when estimating the log-relative risk function, although the LL method can be preferable when there is a linear source of risk. We obtained asymptotic formula for the bias and variance of local linear estimator.

As with all kernel smoothing problems, the choice of bandwidth is critical. Likelihood cross-validation bandwidth selection has proven successful when the target function is the conditional probability, but is much less attractive when we seek to estimate the log-relative risk function. Indeed, the worse performance of the local-linear estimator implemented with the likelihood cross-validation bandwidth is perhaps the most striking outcome of our numerical experiments.

Chapter 5

Estimation of spatio-temporal relative risk function

5.1 Introduction

Abellan, Richardson & Best (2008) have mentioned in the past two decades, that geographic information systems and statistical methods for analyzing spatially referenced data allowed epidemiologists to re-evaluate spatial epidemiology from the perspective of visualizing the trajectory of disease risk across space and time. So there is a growing interest in integrating temporal information in geographic information systems, while advances in computing technologies have made it possible to implement many of the new concepts developed to address temporal problems (Lo & Yeung, 2007). The general aim of this chapter is to display temporal changes in the spatial pattern of disease. As in other chapters, we use kernel density estimation to compute the spatio-temporal estimators.

This chapter has two parts. The first part is organized as follows. In Section 5.2, we extend spatial relative risk function (Bithel, 1990) to a spatio-temporal relative risk function. Spatio-temporal kernel density estimation is defined in the context of multivariate data in Section 5.3. It is usual to observe some observations at the boundary of the region. So we employ the edge correction techniques detailed in Section 5.4. In Section 5.5, we derive formulae for the asymptotic properties of bias and variance of the spatio-temporal kernel estimates and relative risk estimates. The proofs of theorems can be found in the Appendix C. We next discuss the tolerance contours of log relative risk in Section 5.6. After that, we describe two cross-validation bandwidth selectors (LSCV and LCV) in the relative risk estimation and obtain the formulae of the optimal bandwidth combination (spatial and temporal) in Section 5.7. In Section 5.8, we illustrate these estimates using the foot and mouth disease data from the 1967 outbreak.

The second part of this Chapter concerns time derivative relative risk estimation since it is very important to see how the rate of a particular disease risk change with time. Some articles regarding kernel density derivative estimation can be found in the literature. Chacón, Duong and Wand (2010) have worked on the kernel estimators of multivariate density derivative functions using general bandwidth matrix selectors. Borla & Protopopescu (2010) have proposed a nonparametric estimator of a fractional derivative of a distribution function. However, the time derivative of log relative risk estimation has not been examined. So our work in this section is a novel contribution.

We organize the rest of the sections as follows. We define the time derivative of log relative risk function in Section 5.9.1. The asymptotic properties of bias and variance of this estimators are obtained in the Section 5.5. Then we derive the tolerance contours of these time derivative density estimators in Section 5.9.2. We revisit the FMD application to illustrate these derivative density estimators in Section 5.9.3.

5.2 Spatio-temporal relative risk function

The basic format of spatio-temporal data is $(\mathbf{x}_i, t_i)$; $(i = 1, ..., n)$ where each $\mathbf{x}_i = (x_{i1}, x_{i2}) \in \mathcal{R}$ represents the continues variables of spatial locations, and $t_i \in T$ denotes the time of occurrence of an event of interest as an ordered continuous variable. One general approach to analysing spatio-temporal data is to extend existing methods for purely spatial data by considering the time of occurrence as a distinguishing feature, attached to each event. So here, we use spatial relative risk function (Bithel, 1990) to extend in the form of space-time. Conditional on the sample sizes n_1 and n_2, we define

$$r(\mathbf{z}, t) = \frac{f(\mathbf{z}, t)}{g(\mathbf{z}, t)};\ t \in T,\ \mathbf{z} \in \mathcal{R} \tag{5.2.1}$$

to be the spatio-temporal relative risk function for $\mathbf{z} \in \mathcal{R}$ occured at time $t \in T$, where f and g represent respectively trivariate probability density functions of cases and controls at spatial coordinates and time of occurence. The average value of relative risk (r), with respect to the control density g is

$$\begin{aligned} E_{\mathbf{z}\sim g, t\sim g}[r(\mathbf{z},t)] &= \int_{\mathcal{R}}\int_{T} r(\mathbf{z},t)g(\mathbf{z},t)d\mathbf{z}dt \\ &= \int_{\mathcal{R}}\int_{T} \frac{f(\mathbf{z},t)}{g(\mathbf{z},t)}g(\mathbf{z},t)d\mathbf{z}dt \\ &= 1. \end{aligned}$$

This implies that the average value of $r(\mathbf{z},t)$ with respect to the space-time control density over R and T is one. That is, $r(\mathbf{z},t) > 1$ indicates the elevated risk of disease at the point $\mathbf{z}$ as compared to the study area as a whole.

It will usually be the case that the control density can be considered unchanging throughout the observational period. We therefore assume $g(\mathbf{z};t) = g(\mathbf{z})g(t) = g(\mathbf{z})\frac{1}{|T|} = \frac{g(\mathbf{z})}{|T|}$, where T represents the time interval and hence $|T|$ the length thereof. Thus, the relative risk function (5.2.1) becomes

$$r(\mathbf{z};t) = \frac{f(\mathbf{z};t)}{g(\mathbf{z})}|T| \ ; \qquad t \in T, \ \mathbf{z} \in \mathcal{R}. \tag{5.2.2}$$

However, there may be some situations when the disease at-risk varies substantially over time, for e.g. when those that have been infected do not become susceptible after infection. In this way, the size of control population will reduce in size and exhibit more spatial variation over time depending on where previous cases have been. In this case, since time (t) is varying, we can apply the theoretical results, obtained in Chapter 6 by replacing the covariate $(\mathbf{z})$ with time (t) to compute the control density $g(\mathbf{x},t)$. However, in this Chapter, we focus on the prior assumption regarding the control density.

As mentioned in previous chapters we will usually work with risk on the log scale; i.e focus will be on the space-time log-relative risk function:

$$\begin{aligned}\rho(\mathbf{z},t) &= \log[r(\mathbf{z},t)] \\ &= \log[f(\mathbf{z},t)] - \log[g(\mathbf{z})] + \log[|T|] \end{aligned} \tag{5.2.3}$$

Throughout this Chapter as in Chapter 2, we prefer estimating ρ to r. A straightforward estimate of ρ (or r) can be obtained by substituting f and g from trivariate kernel density estimates constructed from case and control data respectively, as we now explain.

5.3 Spatio-temporal kernel density estimation

The trivariate kernel density estimate constructed from independent, identically distributed bivariate spatial observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and univariate temporal data $t_1, \ldots, t_{n_1}$ of size n_1 is given by

$$\hat{f}(\mathbf{z};t) = n_1^{-1} \sum_{i=1}^{n_1} K_h\left(\mathbf{z} - \mathbf{x}_i\right) L_\lambda\left(t - t_i\right), \tag{5.3.1}$$

where $K_h(\mathbf{u}) = h^{-2}K(h^{-1}\mathbf{u})$ and $L_\lambda(t) = \lambda^{-1}L(\lambda^{-1}t)$. In this study as in Chapter 2, the kernel functions K and L are defined using a spherically symmetric probability density function satisfying $\int K(\mathbf{u})d\mathbf{u} = 1$ and a univariate probability density function satisfying $\int L(t)dt = 1$, respectively. The scalars h and λ determine the bandwidths for spatial and temporal components respectively and these bandwidths control the smoothness of respective density estimators. See Wand & Jones (1995) and Duong & Hazelton (2003) for more details.

Since

$$\hat{g}(\mathbf{z}, t) = \frac{\hat{g}(\mathbf{z})}{|T|},$$

the kernel density estimator $\hat{g}$ of g can be expressed as

$$\hat{g}(\mathbf{z}; t) = \frac{n_2^{-1}}{|T|} \sum_{i=n_1+1}^{n_1+n_2} K_{h_g}(\mathbf{z} - \mathbf{x}_i), \tag{5.3.2}$$

where h_g is the smoothing parameter of the kernel density estimate and $\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_1+n_2}$ represents the control observations of size n_2. If we assume that h_f be the bandwidth for the kernel estimate computed from case data, a plug-in estimator of space-time log-relative risk function from (5.2.3) is then given by

$$\hat{\rho}(\mathbf{z}, t) = \log[\hat{f}(\mathbf{z}, t; h_f, \lambda)] - \log[\hat{g}(\mathbf{z}; h_g)] + \log[|T|]. \tag{5.3.3}$$

A major problem we have to overcome is selecting the bandwidths. As there is no prior research in the estimation of the spatio-temporal relative risk function, we use a common bandwidth which is preferred in most applications since there are some substantial complication when dealing with separate bandwidths. The time/spatial scale are entirely arbitrary, so non-equal bandwidths can easily be resolved with appropriate rescaling anyway. That is, $(h_f = h_g = h)$ in deriving $\hat{f}$ and $\hat{g}$.

5.4 Edge correction

As mentioned in Chapter 2 in the context of spatial analysis, edge correction techniques can be applied to avoid the bias near boundaries when estimating the log relative risk function. In the context of space-time, because there may be data points that fall close to the boundary of $\mathcal{R}$, and or to the ends of the observational time

window.

Fixed bandwidth edge corrected spatio-temporal kernel density estimate constructed from case data can be defined as

$$\tilde{f}_{h_f}(\mathbf{z},t) = \frac{\hat{f}_{h_f}(\mathbf{z},t)}{q(\mathbf{z},t)}$$

where q is the explicit edge correction term and

$$q(\mathbf{z},t) = \int_T \int_{\mathcal{R}} K_{h_f}(\mathbf{y}-\mathbf{z}) L_\lambda(s-t) d\mathbf{y} ds \tag{5.4.1}$$

$$= \int_{\mathcal{R}} K_{h_f}(\mathbf{y}-\mathbf{z}) d\mathbf{y} \int_T L_\lambda(s-t) ds. \tag{5.4.2}$$

The integral in the left side of equation (5.4.2) represents the bivariate edge-correction term due to the spatial locations while the right side is a univariate term due to the time component. Regarding the latter, we explain some issues. If we observe all cases in some specific epidemic then we know that there are no cases outside the time interval T, and hence no boundary correction in time is required. If we only observe cases over T and there may be cases before or after, and then boundary correction in time is required. The former is appropriate for the FMD data, for example. When we ignore the boundary correction in time, $q(\mathbf{z},t)$ becomes $q(\mathbf{z})$. From equation (5.2.2), the boundary corrected relative risk estimate can then be revised as

$$\begin{aligned} \tilde{r}(\mathbf{z};t) &= \frac{\tilde{f}(\mathbf{z};t)\ |T|}{\tilde{g}(\mathbf{z})}; \qquad t \in T,\ \mathbf{z} \in \mathcal{R}. \\ &= \frac{\hat{f}_{h_f}(\mathbf{z},t)\ |T| q_{h_g}(\mathbf{z})}{q_{h_f}(\mathbf{z}) \hat{g}(\mathbf{z})} \\ &= \hat{r}(\mathbf{z};t) \frac{q_{h_g}(\mathbf{z})}{q_{h_f}(\mathbf{z})} \end{aligned} \tag{5.4.3}$$

The top and bottom edge correction terms in equation (5.4.3) cancels out if common

spatial bandwidths ($h_f = h_g = h$) are used for case and control densities. So the equation (5.4.3) then reduces to $\hat{r}(\mathbf{z}, t)$.

5.5 Asymptotic properties of spatio-temporal relative risk estimators

In this Section, we derive the asymptotic properties of bias and variance of space-time kernel density estimates, i.e. $\hat{f}(\mathbf{z}, t)$, and log RR estimates $\hat{\rho}(\mathbf{z}, t)$ for $\mathbf{z}$ lying in the interior of the study region $\mathcal{R}$. We assume common bandwidths for case and control data in the estimation.

Lemma 5.5.1. *If $\boldsymbol{z}$ be an interior point of the region $\mathcal{R}$ then at a particular time t,*

$$Bias[\hat{f}(\boldsymbol{z}, t)] = \frac{1}{2}h^2\mu_2(K)\nabla_z^2 f(\boldsymbol{z}, t) + \frac{1}{2}\lambda^2\mu_2(L)\nabla_t^2 f(\boldsymbol{z}, t) + o(h^2 + \lambda^2) \tag{5.5.1}$$

and

$$Var[\hat{f}(\boldsymbol{z}, t)] = \frac{f(\boldsymbol{z}, t)R(K)R(L)}{n_1 h^2 \lambda} + o(n_1^{-1}h^{-2}\lambda^{-1}). \tag{5.5.2}$$

Here ∇, $\mu_2(K)$, K and $R(K)$ have already been defined in Chapter 4, while $\mu_2(L) = \int s^2 L(s)ds$ and $R(L) = \int L(s)^2 ds$.

The proof of this lemma follows from the standard results for multivariate density estimation (e.g. Wand & Jones, 1995).

Theorem 5.5.2. *If $\boldsymbol{z}$ be an interior point of $\mathcal{R}$ then*

$$Bias[\hat{\rho}(\boldsymbol{z}, t)] = \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla_z^2 f(\boldsymbol{z}, t) + \lambda^2\mu_2(L)\nabla_t^2 f(\boldsymbol{z}, t)}{f(\boldsymbol{z}, t)} - \frac{h^2\mu_2(K)\nabla_z^2 g(\boldsymbol{z})}{g(\boldsymbol{z})}\right] + o(h^2 + \lambda^2) \tag{5.5.3}$$

and

$$Var[\hat{\rho}(\boldsymbol{z},t)] = \frac{R(K)R(L)f(\boldsymbol{z},t)^{-1}}{n_1 h^2 \lambda} + \frac{R(K)g(\boldsymbol{z})^{-1}}{n_2 h^2} + o(n_1^{-1}h^{-2}\lambda^{-1} + n_2^{-1}h^{-2}) \quad (5.5.4)$$

at time t.

The proof of this theorem can be seen in the Appendix C.1. According to these asymptotic properties, we should avoid the possibility of being $g = 0$ and $f = 0$. As in Chapters 2 and 4, we assume that $f(\mathbf{z}, t) > \epsilon$ and $g(\mathbf{z}) > \epsilon$ for all $\mathbf{z} \in \mathcal{R}$ and $t \in T$, where ϵ is assumed known.

Here the variance formulae (5.5.4) of $\hat{\rho}$ is important since we use it to compute z-test statistics for tolerance contours.

It will sometimes be the case that time is recorded at a relatively coarse level, so that there are only a modest number of unique values for t. When that occurs, the time bandwidth (λ) will be larger than spatial bandwidth, h. In that case, the second term in (5.5.1) and (5.5.3) will be the dominant element in both bias expressions.

5.6 Tolerance contours of $\hat{\rho}$

Tolerance counters are often appropriate the analysis of relative risk function to be able to spot statistically significant variations in the risk itself. See Kelsall & Diggle (1995a) in the context of spatial analysis and the discussion in Chapter 2. So we extend the work in Chapter 2 regarding tolerance contours into the context of space-time. We aim to identify the subregions which produce the elevated risk surface by

considering the null hypothesis

$$H_0 : \rho(\mathbf{z}, t) = 0 \;\; (i.e.\; r(\mathbf{z}, t) = 1)$$

against the alternative hypothesis

$$H_1 : \rho(\mathbf{z}, t) > 0 \;\; (i.e.\; r(\mathbf{z}, t) > 1).$$

As mentioned in Chapter 2, this produces a p-values surface $p(\mathbf{z}, t)$ over the region $\mathcal{R}$ at time t. If we compute the contour relevant to $p = 5\%$, then all space-time points within this contour (i.e. with $p(\mathbf{z}; t) \leq 5\%$) can be considered to have significantly elevated risk. Here we therefore suggest that the p-values for constructing space-time tolerance contours for the density ratio estimator be obtained by using a normal approximation to the sampling distribution of $\hat{\rho}(\mathbf{z}; t)$.

Under the null hypothesis H_0, $r(\mathbf{z}, t) = 1$, i.e., $f(\mathbf{z}, t) = g(\mathbf{z})\frac{1}{|T|}$. Then the asymptotic bias of $\hat{\rho}(\mathbf{z}; t)$ in equation (5.5.3) becomes zero and the asymptotic variance from equation (5.5.4) reduces to

$$\begin{aligned} \mathrm{Var}\hat{\rho}(\mathbf{z}, t) &\approx \frac{R(K)R(L)f(\mathbf{z}; t)^{-1}}{n_1 h^2 \lambda} + \frac{R(K)g(\mathbf{z})^{-1}}{n_2 h^2} \\ &\approx \frac{R(K)R(L)|T|g(\mathbf{z})^{-1}}{n_1 h^2 \lambda} + \frac{R(K)g(\mathbf{z})^{-1}}{n_2 h^2} \\ &\approx \frac{R(K)g(\mathbf{z})^{-1}}{h^2}\left[|T|\frac{R(L)}{n_1 \lambda} + \frac{1}{n_2}\right]. \end{aligned}$$

If $g(\mathbf{z})$ is unknown, we use pooled estimate $\hat{g}_\rho(\mathbf{z})$, constructed from all data consist of both cases and controls. Hence an appropriate z-test statistic at the point $\mathbf{z}$ at time t can be obtained by using

$$Z(\mathbf{z}; t) \;=\; \frac{\hat{\rho}(\mathbf{z}; t)}{SE\{\hat{\rho}(\mathbf{z}; t)\}} \;;\; \text{under } H_0,$$

where

$$SE\{\hat{\rho}(\mathbf{z},t)\} = \sqrt{\frac{R(K)\hat{g}_\rho(\mathbf{z})^{-1}}{h^2}\left[|T|\frac{R(L)}{n_1\lambda}+\frac{1}{n_2}\right]}.$$

5.7 Bandwidth selection

The choice of smoothing parameter is a crucial element in spatial relative risk estimation (see Chapter 3). This selection problem is more difficult in the context of space-time since the temporal bandwidth needs to be incorporated in addition to the usual spatial bandwidth. One can interactively choose suitable bandwidths by looking at different bandwidths. Automatic bandwidth estimators can also be applied in this case.

The MISE of log RR estimate at a point $\mathbf{z}$ at time t is given by

$$\text{MISE}(h,\lambda) = \mathbb{E}\left[\int_T\int_{\mathcal{R}}[\hat{\rho}(\mathbf{z},t)-\rho(\mathbf{z},t)]^2 d\mathbf{z}dt\right]$$

where h is the common case and control spatial bandwidth and λ is the temporal bandwidth. Using the standard properties of mean and variance,

$$\text{MISE}(h,\lambda) = \int_T\int_{\mathcal{R}}\left[\text{Bias}\{\widehat{\rho}(\mathbf{z},t)\}\right]^2 d\mathbf{z}dt + \int_T\int_{\mathcal{R}}\text{Var}\{\widehat{\rho}(\mathbf{z},t)\}d\mathbf{z}dt. \tag{5.7.1}$$

Using the equations (5.5.3) and (5.5.4) for the mean and variance of $\hat{\rho}$, we reach to an asymptotic approximation to the MISE, given by

$$\text{MISE}(h,\lambda) \approx \frac{1}{4}\int_T\int_{\mathcal{R}}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}f(\mathbf{z},t)+\lambda^2\mu_2(L)\frac{\partial^2}{\partial t^2}f(\mathbf{z},t)}{f(\mathbf{z},t)} - \frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})}\right]^2 d\mathbf{z}dt$$
$$+\int_T\int_{\mathcal{R}}\left[\frac{R(K)R(L)f(\mathbf{z},t)^{-1}}{n_1h^2\lambda}+\frac{R(K)g(\mathbf{z})^{-1}}{n_2h^2}\right]d\mathbf{z}dt.$$

Here our aim is to obtain the optimal bandwidth combination (h, λ) by minimizing MISE. So we get the optimal bandwidths,

$$\widehat{(h,\lambda)}_{\text{MISE}} = \arg\min_{h,\lambda}[\text{MISE}].$$

In principle the LSCV and LCV bandwidth selectors can be used in space-time relative risk estimation. These two methods are briefly described in the context of space-time in the following two subsections.

5.7.1 Least squares cross validation

The space-time log relative risk function from equation (5.2.3) is,

$$\rho(\mathbf{z}, t) = \log\left[\frac{f(\mathbf{z}, t)}{g(\mathbf{z})}|T|\right]; \quad t \in T, \mathbf{z} \in \mathcal{R}. \tag{5.7.2}$$

We obtain the LSCV formula for spatio-temporal case, derived from Chapter 3 equation (3.3.3) as follows:

$$\begin{aligned}\widehat{LSCV}(h,\lambda) &= -\int\int\{\hat{\rho}_{h,\lambda}(\mathbf{x},t)\}^2 d\mathbf{x}dt - 2n_1^{-1}\sum_{i=1}^{n_1}\log\left[\frac{(\hat{f}^{-i}_{h,\lambda}(\mathbf{x}_i,t_i))|T|}{\hat{g}_h(\mathbf{x}_i)}\right]\{\hat{f}^{-i}(\mathbf{x}_i,t_i)\}^{-1} \\ &\quad +2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2}\log\left[\frac{(\hat{f}_{h,\lambda}(\mathbf{x}_j,t_j))|T|}{\hat{g}_h^{-j}(\mathbf{x}_j)}\right]\left[\frac{\hat{g}^{-j}(\mathbf{x}_j)}{|T|}\right]^{-1}. \end{aligned}\tag{5.7.3}$$

When $\mathbf{x}$ lies at the boundary of the region, the equation (5.7.3) needs to be revised as

$$\begin{aligned}\widehat{LSCV}(h,\lambda) &= -\int\int\{\tilde{\rho}_{h,\lambda}(\mathbf{x},t)\}^2 d\mathbf{x}dt - 2n_1^{-1}\sum_{i=1}^{n_1}\log\left[\frac{(\tilde{f}^{-i}_{h,\lambda}(\mathbf{x}_i,t_i))|T|}{\tilde{g}_h(\mathbf{x}_i)}\right]\{\tilde{f}^{-i}(\mathbf{x}_i,t_i)\}^{-1} + \\ &\quad 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2}\log\left[\frac{(\tilde{f}_{h,\lambda}(\mathbf{x}_j,t_j))|T|}{\tilde{g}_h^{-j}(\mathbf{x}_j)}\right]\left[\frac{\tilde{g}^{-j}(\mathbf{x}_j)}{|T|}\right]^{-1}. \end{aligned}\tag{5.7.4}$$

As mentioned in Section 5.4, under common case and control bandwidths, $\tilde{\rho} = \hat{\rho}$, so the first term of equation (5.7.3) remains as it is. The second term of equation (5.7.4) is reduced to the second term in equation (5.7.3) multiplied by $q(\mathbf{z})$. Similarly, the third term of equation. (5.7.4) is reduced to the third term in equation (5.7.3) multiplied by $q(\mathbf{z})$, where

$$q(\mathbf{z}) = \int_{\mathcal{R}} \frac{1}{h^2} K\left(\frac{\mathbf{x}-\mathbf{z}}{h} d\mathbf{x}\right).$$

The minimization of the estimate $\widehat{LSCV}$ over h and λ produces the LSCV optimal combination of the spatial and temporal bandwidths.

5.7.2 Likelihood cross-validation

We present in this section an alternative method, LCV for choosing the optimal bandwidth combination. See Clark and Lawson (2004) for spatial analysis.

Let f and g be spatio-temporal bivariate density functions constructed from case and control data sets respectively. This is an extension of the LCV discussion in the estimation of spatial relative risk in Section 3.3.2 to the spatio-temporal context. Let us define the regression function $p(\mathbf{z}, t) = E[Y|\mathbf{X} = \mathbf{z}, T = t]$, in which y is the binary response function with $y = 1$ for case data and $y = 0$ for control data. That is, $p(\mathbf{z}, t) = P[Y = 1|\mathbf{X} = \mathbf{z}, T = t]$. Then,

$$\begin{aligned} p(\mathbf{z}, t) &= \frac{\pi f(\mathbf{z}, t)}{(1-\pi)g(\mathbf{z}, t) + \pi f(\mathbf{z}, t)}; \quad \pi = \frac{n_1}{n_1 + n_2} \\ &= \frac{n_1 f(\mathbf{z}, t)}{n_2 g(\mathbf{z})/|T| + n_1 f(\mathbf{z}, t)} \end{aligned}$$

To choose spatial (h) and temporal(λ) optimal bandwidths, we use likelihood cross validation (see Clark and Lawson (2004) for spatial analysis). We define the LCV function

$$\begin{aligned}
\widehat{LCV}(h,\lambda) &= \log\left[\prod_{i=1}^{n}\{\hat{p}^{-i}(\mathbf{x}_i,t_i;h,\lambda)\}^{y_i}\{1-\hat{p}^{-i}(\mathbf{x}_i,t_i;h,\lambda)\}^{1-y_i}\right] \\
&= \sum_{i=1}^{n}\left[y_i\log[\hat{p}^{-i}(\mathbf{x}_i,t_i;h,\lambda)]+(1-y_i)\log[1-\hat{p}^{-i}(\mathbf{x}_i,t_i;h,\lambda)]\right] \\
&= \sum_{i=1}^{n_1}\left[y_i\log[\hat{p}_1^{-i}(\mathbf{x}_i,t_i;h,\lambda)]+(1-y_i)\log[1-\hat{p}_1^{-i}(\mathbf{x}_i,t_i;h,\lambda)]\right]+ \\
&\quad \sum_{i=n_1+1}^{n_1+n_2}\left[y_i\log[\hat{p}_2^{-i}(\mathbf{x}_i;h)]+(1-y_i)\log[1-\hat{p}_2^{-i}(\mathbf{x}_i;h)]\right].
\end{aligned}$$

Here,

$$\hat{p}_1^{-i}(\mathbf{x}_i,t_i;h,\lambda)=\frac{n_1\hat{f}^{-i}(\mathbf{x}_i,t_i,h,\lambda)|T|}{n_2\hat{g}(\mathbf{x}_i;h)+n_1\hat{f}^{-i}(\mathbf{x}_i,t_i,h,\lambda)|T|}$$

and

$$\hat{p}_2^{-i}(\mathbf{x}_i;h)=\frac{n_2\hat{g}^{-i}(\mathbf{x}_i;h)}{n_1\hat{f}(\mathbf{x}_i;h)+n_2\hat{g}^{-i}(\mathbf{x}_i;h)}.$$

By substituting $\hat{p}_1^{-i}$ and $\hat{p}_2^{-i}$,

$$\widehat{LCV}(h,\lambda)=\sum_{i=1}^{n_1}\log\left[\frac{n_1\hat{f}^{-i}(\mathbf{x}_i,t_i)}{\frac{n_2\hat{g}(\mathbf{x}_i)}{|T|}+n_1\hat{f}^{-i}(\mathbf{x}_i,t_i)}\right]+\sum_{i=n_1+1}^{n_1+n_2}\log\left[\frac{n_2\hat{g}^{-i}(\mathbf{x}_i)}{n_2\hat{g}^{-i}(\mathbf{x}_i)+n_1\hat{f}(\mathbf{x}_i)}\right].$$

Here $\hat{f}^{-i}$ and $\hat{g}^{-i}$ represent the kernel density estimates constructed from all data points except i^{th} datum. $\widehat{LCV}$ needs to be maximized to obtain the optimal bandwidth combination.

5.8 Real application: FMD of 1967 outbreak

In this section, we revisit the data from the 1967 outbreak of FMD in Cumbria, UK. We have already examined these data in Section 4.8.3, but only with respect to the spatial locations of cases and controls. In this Section, in addition to spatial points we include time of infection (measured in days).

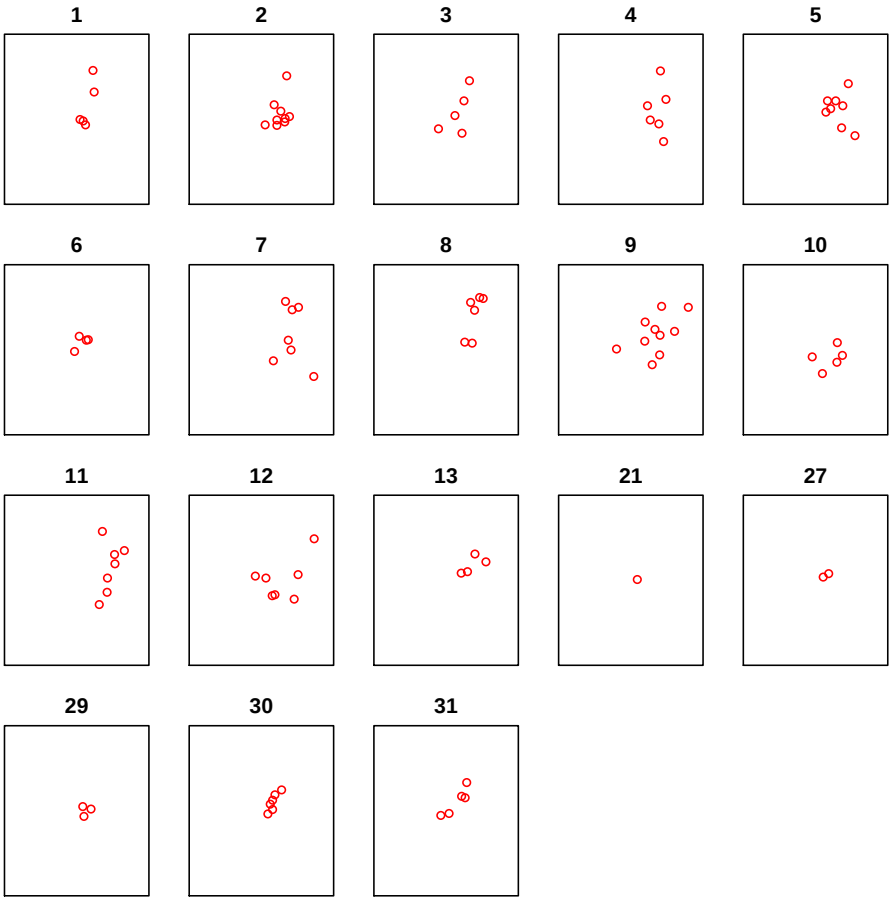

Figure 5.1: The distribution of FMD case data. Day is shown in the title.

According to Figure 5.1, there are clusters of case data, found in the middle and eastern parts of the region. Note that there were no cases on days 14-20, 22-26 and

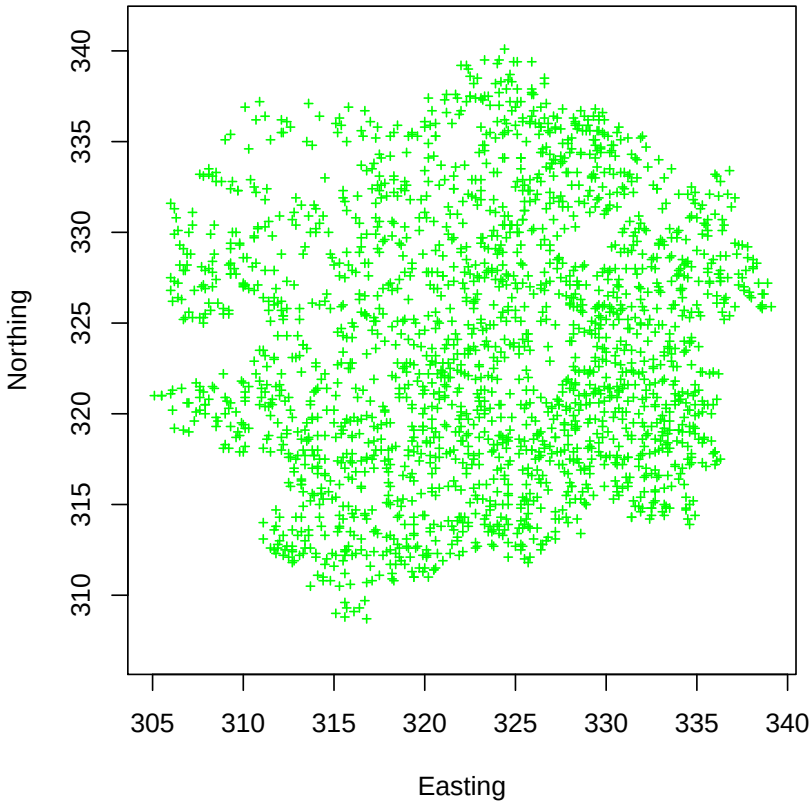

Figure 5.2: Control data distribution of FMD.

28. Figure 5.2 depicts the distribution of 2129 controls, which represent uninfected farms at risk of the disease. Our primary objective here is to examine the temporal changes in the spatial pattern of FMD disease.

We compute log relative risk estimates of FMD using the density ratio method in the context of space-time. Then we highlight the farms which correspond to significantly

elevated risk using the 95% tolerance contours. Initially, we choose subjective bandwidths, $\hat{h} = 1.2$ for spatial component while $\hat{\lambda} = 1.7$ for temporal.

The resultant contour plots representing the log relative risk estimates are given in Figure 5.3. These four plots depict the spatio-temporal relative risk surface of FMD on days 1, 6, 16 and 24. The overall impression is that the disease risk has become more highly dispersed over the region between days 1 and 24, with the highest area of risk moving eastward a little followed by centralizing again at the end of the period.

Figure 5.4 shows the log relative risk estimates computed on days 1, 6, 16 and 24 as previously, but using LSCV bandwidths 0.83 and 2.15 for spatial and temporal components respectively. These results seem comparable with the results from using the subjective bandwidths and this produces an under-smoothed estimate.

5.9 Time derivative relative risk estimation

Estimating time derivatives of the spatio-temporal log relative risk surface is very important in spatio-temporal analysis, since it can be used to measure the rate of change of the log relative risk of a disease over time. We build on published results on kernel estimation of density derivatives (e.g. Chacón, Duong and Wand (2011). The discussion of this section provides novel contributions with the introduction to time derivatives of the relative risk function and also its asymptotic properties and then the computation of the 5% tolerance contours to identify areas where the relative risk is changing.

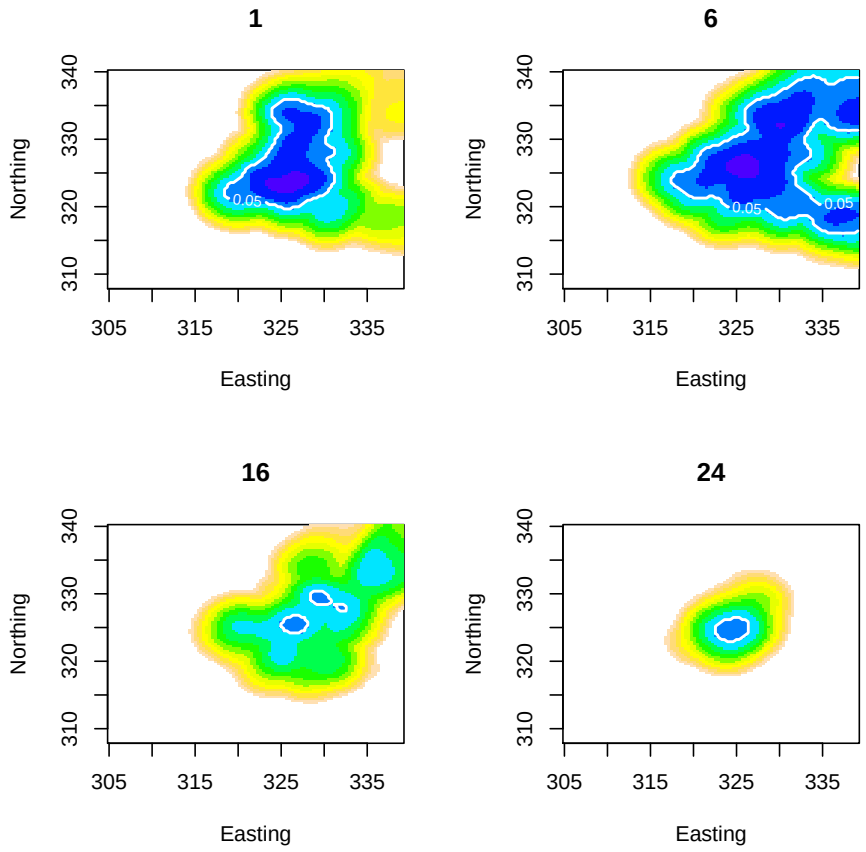

Figure 5.3: Log relative risk estimates(ρ) for FMD data in day 1, 6, 16 and 24. Subjective bandwidth is used. 5% tolerance contours are displayed in white.

5.9.1 Time derivative density estimation

Bhattacharya (1967) has worked on the estimation of a univariate probability density function and its derivatives and later by Schuster (1969). Singh (1987) studied on the MISE of kernel estimates of a density and its derivatives. Hädle, Marron and Wand (1990) looked for the bandwidth selectors for density derivatives. Jones (1994) generalized the multivariate kernel density derivative estimation.

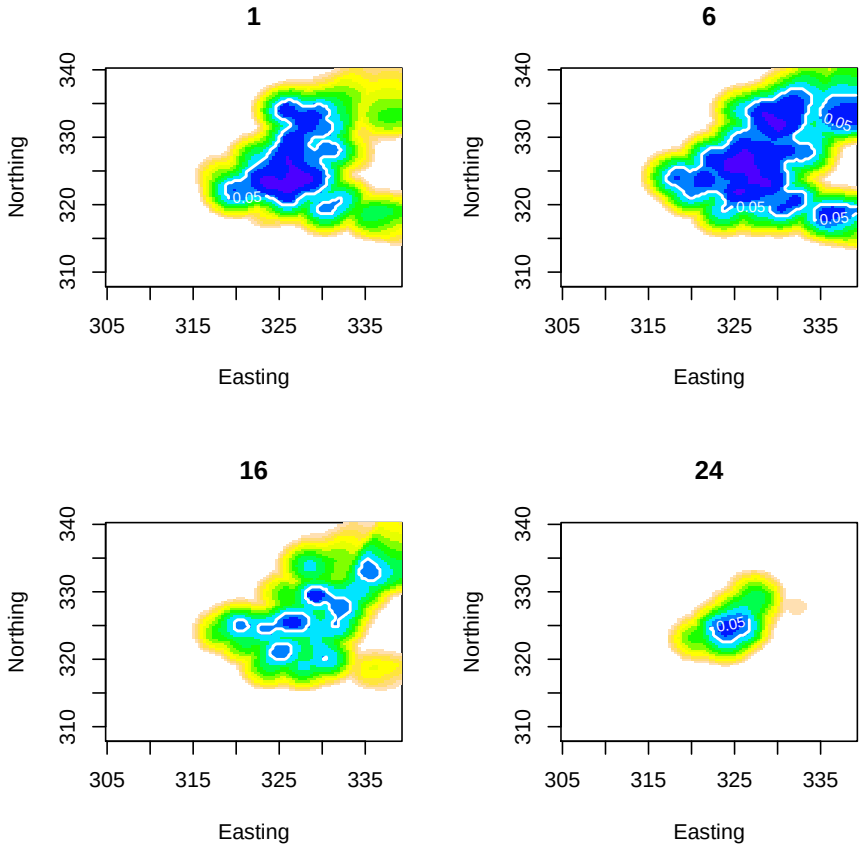

Figure 5.4: Log relative risk estimates (ρ) for FMD data in day 1, 6, 16 and 24. LSCV bandwidth is used. 5% tolerance contours are displayed in white.

The log relative risk time derivative density function is defined by $\frac{\partial}{\partial t}\rho(\mathbf{z};t)$, where $\mathbf{z} \in \mathcal{R}$, $t \in T$.

From equation (5.3.3),

$$\rho(\mathbf{z};t) \quad = \quad \log f(\mathbf{z};t) + \log|T| - \log g(\mathbf{z})$$

as the second and third terms are independent of t, we get

$$\frac{\partial}{\partial t}\rho(\mathbf{z};t) = \frac{\frac{\partial}{\partial t}f(\mathbf{z};t)}{f(\mathbf{z};t)}.$$

That is, time derivatives of log relative risk function is equal to the ratio of time derivatives of case density to the density itself. We can derive this explicitly by plugging in a kernel density estimate of f and its time derivative, to give

$$\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t) = \frac{\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)}{\hat{f}(\mathbf{z};t)}.$$

We now derive explicit expressions for the estimators and constants where K and L are bivariate and univariate Gaussian kernels respectively.

$$\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t) = \frac{1}{n_1 h^2 \lambda}\sum_{i=1}^{n_1} K\left(\frac{\mathbf{z}-\mathbf{x}_i}{h}\right)\frac{\partial}{\partial t}L\left(\frac{t-t_i}{\lambda}\right) \tag{5.9.1}$$

with

$$\begin{aligned}\frac{\partial}{\partial t}L\left(\frac{t-t_i}{\lambda}\right) &= \frac{\partial}{\partial t}\left[\frac{1}{\sqrt{2\pi}\lambda}\exp[-\frac{1}{2\lambda^2}(t-t_i)^2]\right] \\ &= -\frac{(t-t_i)}{\sqrt{2\pi}\lambda^3}\exp[-\frac{1}{2\lambda^2}(t-t_i)^2] \\ &= -\frac{(t-t_i)}{\lambda^2}L\left(\frac{t-t_i}{\lambda}\right)\end{aligned}$$

From equation (5.9.1),

$$\begin{aligned}\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t) &= -\frac{1}{n_1 h^2 \lambda}\sum_{i=1}^{n_1} K\left(\frac{\mathbf{z}-\mathbf{x}_i}{h}\right)\frac{(t-t_i)}{\lambda^2}L\left(\frac{t-t_i}{\lambda}\right) \\ &= -\frac{1}{n_1 \lambda^2}\sum_{i=1}^{n_1} K_h(\mathbf{z}-\mathbf{x}_i)(t-t_i)L_\lambda(t-t_i).\end{aligned}$$

Theorem 5.9.1. *If $\boldsymbol{z}$ be an interior point of $\mathcal{R}$ then*

$$Bias\left[\frac{\partial}{\partial t}\hat{\rho}(\boldsymbol{z};t)\right] = \frac{R(K)R(L)\frac{\partial}{\partial t}f(\boldsymbol{z};t)}{2n_1 h^2 \lambda\ [f(\boldsymbol{z},t)]^2} + o(n_1^{-1}h^{-2}\lambda^{-1}) \tag{5.9.2}$$

and

$$Var\left[\frac{\partial}{\partial t}\hat{\rho}(\boldsymbol{z};t)\right] = \frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\boldsymbol{z};t)} + o\left(\frac{1}{n_1 h^2 \lambda} + \frac{1}{n_1 h^2 \lambda^3}\right). \tag{5.9.3}$$

at time t.

The proof of the above theorem can be found in the Appendix C.2. The term $\frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\mathbf{z};t)}$ of equation (5.9.3) dominates asymptotically because it relates to the estimation of a density derivative, which is inherently a more difficult estimation problem than estimating the density itself.

5.9.2 Tolerance contours of $\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)$

Tolerance contours are discussed in this section to distinguish significant features of the time derivative of the log-relative risk estimator. A particular aim is to examine which parts of the region are seeing real changes in disease risk with time.

As in Section 5.6, we compute 5% tolerance contours. That is, the contours within the p-value, $p = 0.05$, then all points within this contour (i.e. with $p(\mathbf{z};t) \leq 0.05$) can be considered to have areas with significant change of risk. We therefore consider p-values for spatio-temporal tolerance contours by using a normal approximation to the sampling distribution of $\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)$.

We will test the null hypothesis

$$H_0 \ : \ \frac{\partial}{\partial t}\rho(\mathbf{z};t) \ = 0$$

against the alternative hypotheses

$$H_1 \ : \ \frac{\partial}{\partial t}\rho(\mathbf{z};t) \ > 0$$

and

$$H_2 \ : \ \frac{\partial}{\partial t}\rho(\mathbf{z};t) \ < 0.$$

However throughout this chapter, we focus on the alternative H_1 since it is most important. Under H_0,

$$\text{Var}\left[\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)\right] = \frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\mathbf{z};t)}.$$

Since the bias of $\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)$ is zero under the null hypothesis H_0, an appropriate z-test statistic at the point $\mathbf{z}$ at time t is given by

$$Z(\mathbf{z};t) \ = \ \frac{\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)}{SE\{\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)\}} \ ; \text{ under } H_0,$$

where

$$SE\left[\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)\right] = \sqrt{\frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\mathbf{z};t)}}.$$

For large n_1, $\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)$ will be normally distributed with zero mean and $\frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\mathbf{z};t)}$ variance. That is,

$$\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t) \sim N\left(0; \frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\mathbf{z};t)}\right) ; \text{ under } H_0$$

Therefore the test statistic at the point $\mathbf{z}$ at time t is,

$$Z(\mathbf{z};t) = \frac{\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)}{\sqrt{\frac{R(K)R(L^{'})}{n_1 h^2 \lambda^3 f(\mathbf{z};t)}}} \ \sim \ N(0;1) \tag{5.9.4}$$

In practice, we replace $f(\mathbf{z},t)$ in the denominator of the test statistic by a pilot estimate thereof. Here

$$\begin{aligned} R(K)R(L^{'}) &= \int (K(\mathbf{u}))^2 \, d\mathbf{u} \int \left(L^{'}(s)\right)^2 ds \\ &= -\int \left(\frac{1}{2\pi} e^{-(u_1^2+u_2^2)/2}\right)^2 du_1 du_2 \int \left(\frac{s}{\sqrt{2\pi}} e^{-\frac{s^2}{2}}\right)^2 ds \\ &= \frac{1}{8\pi^3} \int e^{-u_1^2} e^{-u_2^2} du_1 du_2 \int s^2 e^{-s^2} ds \\ &= \frac{1}{8\pi^3} \int \int e^{-u_1^2} du_1 \int e^{-u_2^2} du_2 \int s^2 e^{-s^2} ds \\ &= \frac{1}{8\pi^3} \sqrt{\pi} \sqrt{\pi} \frac{\sqrt{\pi}}{4} \\ &= \frac{1}{32\pi^{3/2}}. \end{aligned}$$

5.9.3 Revisit to FMD application

This is a revisit to FMD data set. We produce time derivative density estimates of log relative risk function $(\frac{\partial}{\partial t}\hat{\rho})$ in this Section. Figures 5.5, 5.6, 5.7, 5.8 and 5.9 show the derivative density estimates of FMD produced within the rectangular window [305, 339.1] $\times$ [308, 340.1]. The degree of change of risk is represented by the shaded region. The solid contour white lines depict the 5% tolerance contours, which give the regions of significant change of risk. Also the case locations are scattered in red in the plot. In this work, for simplicity, we use spatial bandwidth, $h = 1.2$ and temporal bandwidth $\lambda = 1.7$ subjectively.

According to the results displayed in these Figures, we make the following brief interpretation. The rate of change in disease risk in Figure 5.5 from day 1 to day 9 shows a decreasing trend though displaying some tiny clusters. On day 10, no cases

of increasing risk can be seen, reflecting the fact that there are no more cases until day 20. According to Figure 5.1, there is a case data on day 21 and further cases on day 27, 29, 30 and 31. Therefore, we can see that there is some significant areas of the change of risk, starting at day 23. After day 31, there are no further cases found and hence the rate of disease risk decreases.

We also use LSCV method to compute spatial bandwidth, $h = 0.83$ and temporal

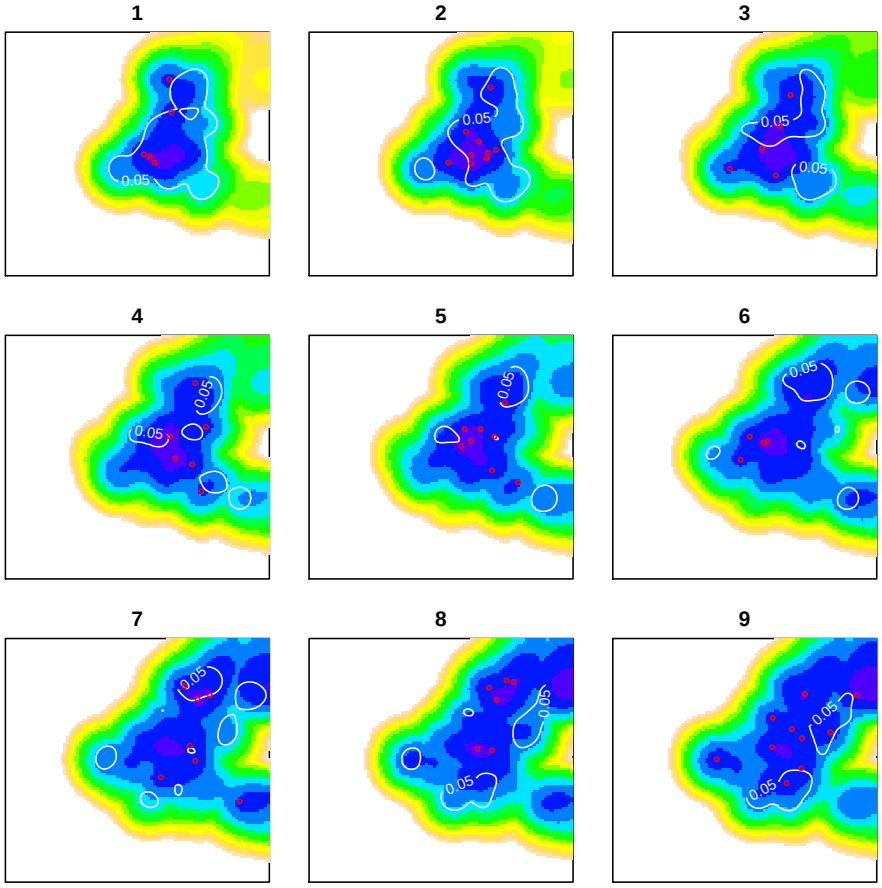

Figure 5.5: Derivative density estimation for days 1-9. White lines indicate 95% tolerance contours. Case data are scattered in red.

bandwidth, $\lambda = 2.15$. The resultant plots are displayed in Figures 5.10, 5.11, 5.12, 5.13 and 5.14, produced within the rectangular window $[305, 339.1] \times [308, 340.1]$.

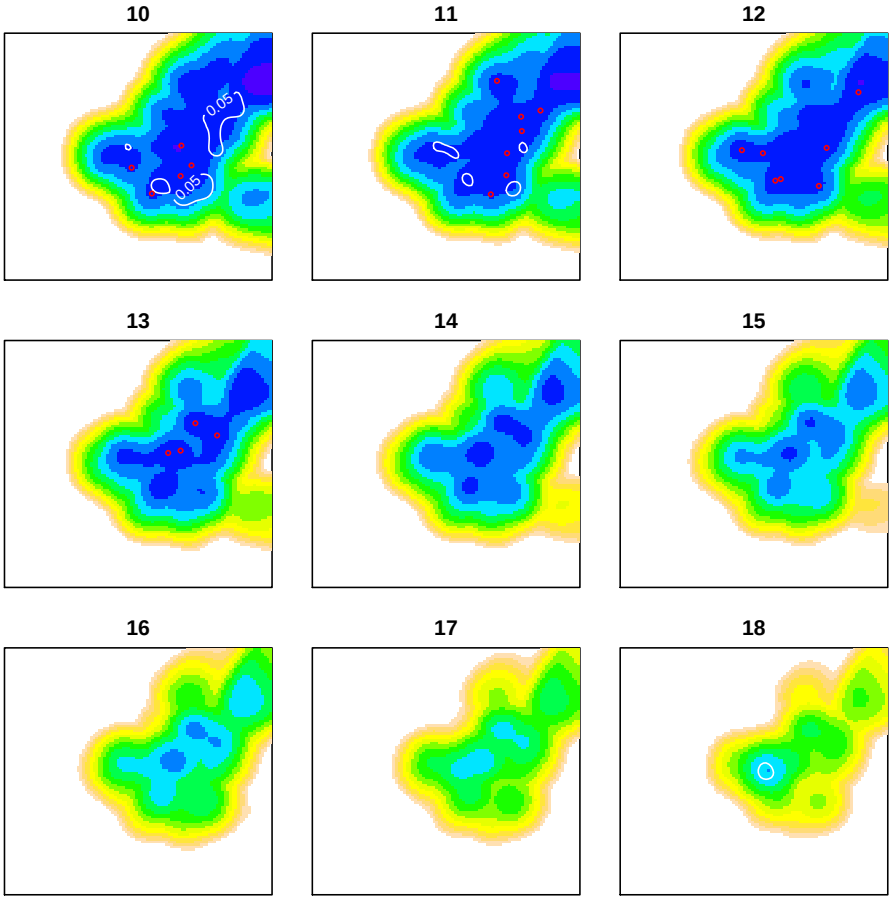

Figure 5.6: Derivative density estimation for days 10-18. White lines indicate 95% tolerance contours. Case data are scattered in red.

Shaded regions represent the degree of change of risk. The 5% tolerance contours are given as solid white contours. Cases can be seen in these plots in red. These resultant plots are very comparable to those with subjective bandwidth. So here, the interpretation is agreeable with the previous case. Also the estimates produced using LSCV looks low smoothed compared to those using subjective bandwidth.

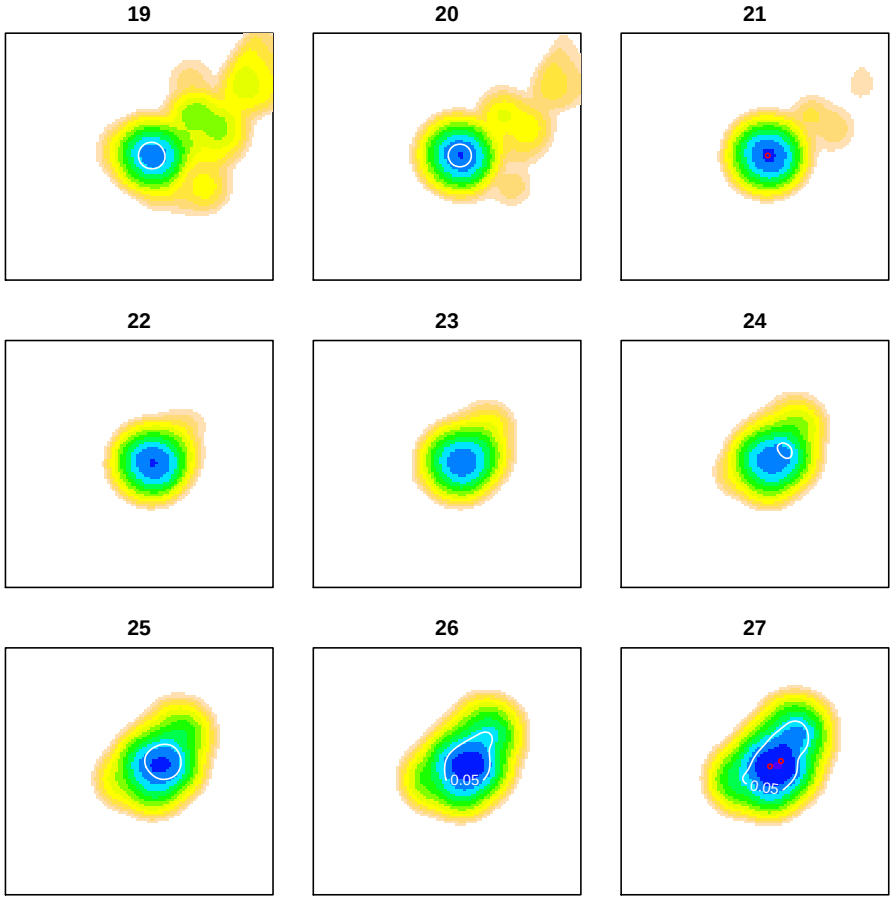

Figure 5.7: Derivative density estimation for days 19-27. White lines indicate 95% tolerance contours. Case data are scattered in red.

5.10 Conclusion

Modelling spatio-temporal patterns of relative risk should be researched since it is an important area and also there is not any previous work in the literature. So in this chapter, we explore the use of kernel density estimation method to identify spatio-temporal patterns of relative risk surfaces from case-control data. Then we obtain the asymptotic properties of bias and variance of log relative risk estimates and its time derivative density estimates. Also we looked at the tolerance contours in order

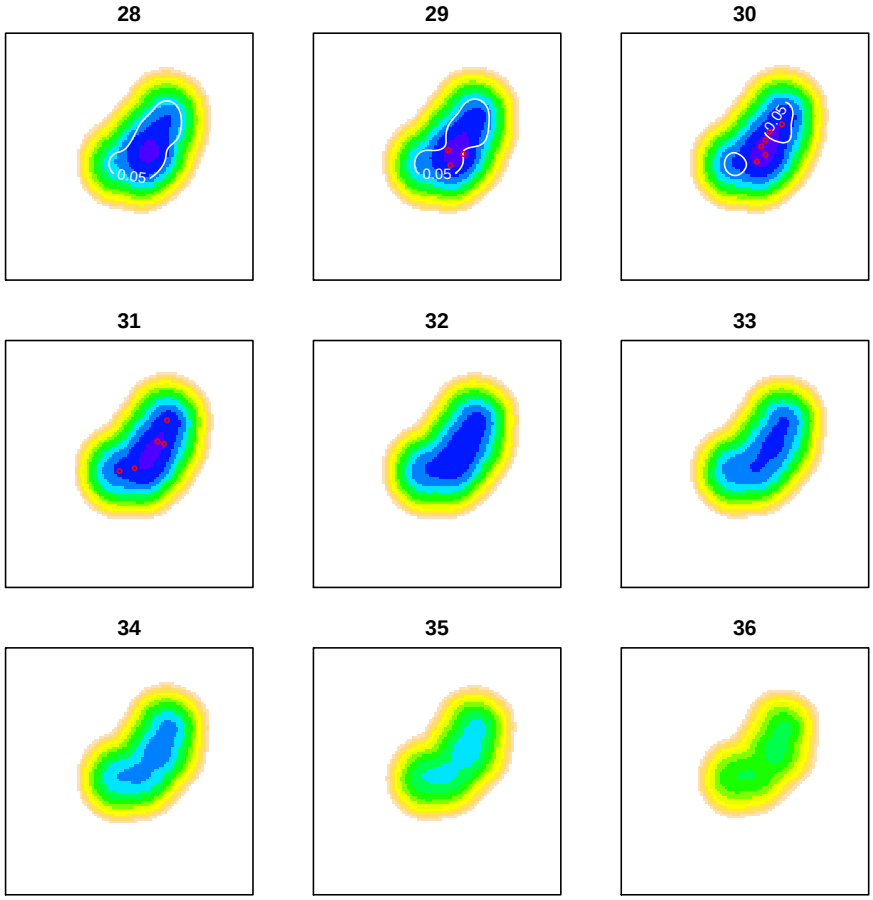

Figure 5.8: Derivative density estimation for days 28-36. White lines indicate 95% tolerance contours. Case data are scattered in red.

to detect the regions of significantly change of disease risk. These relative risk and time derivative density estimates are illustrated using the FMD data set in 1967 outbreak. Spatial and temporal bandwidths have been chosen subjectively and LSCV to compute these estimates.

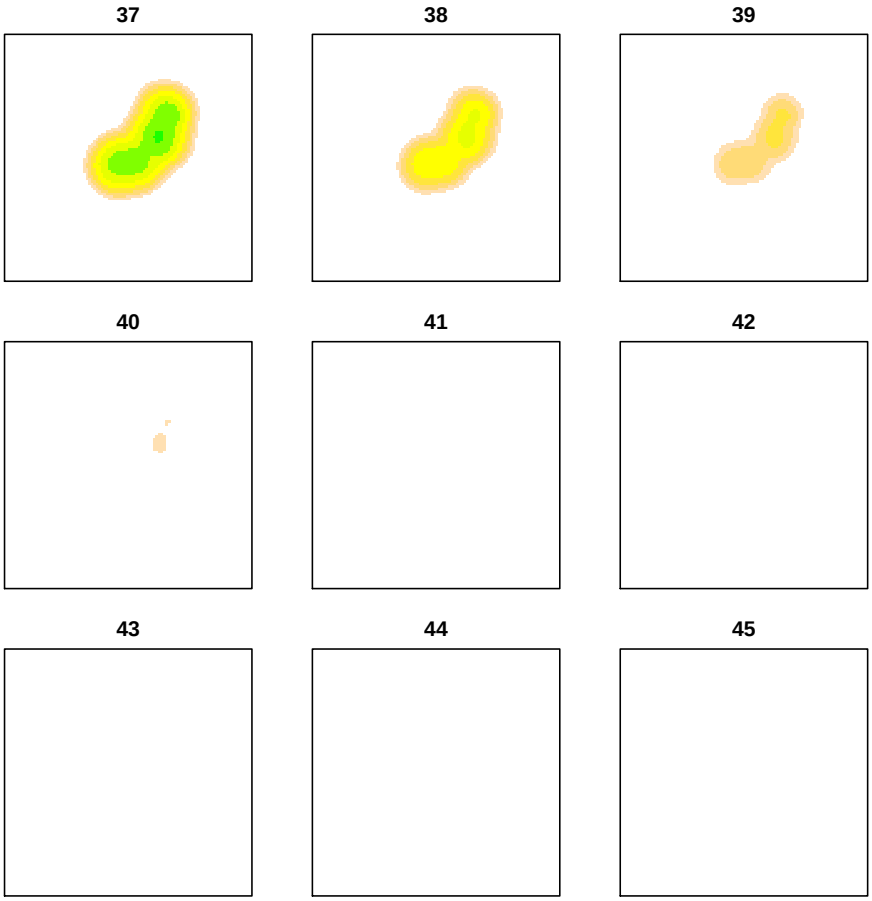

Figure 5.9: Derivative density estimation for days 37-45. White lines indicate 95% tolerance contours. Case data are scattered in red.

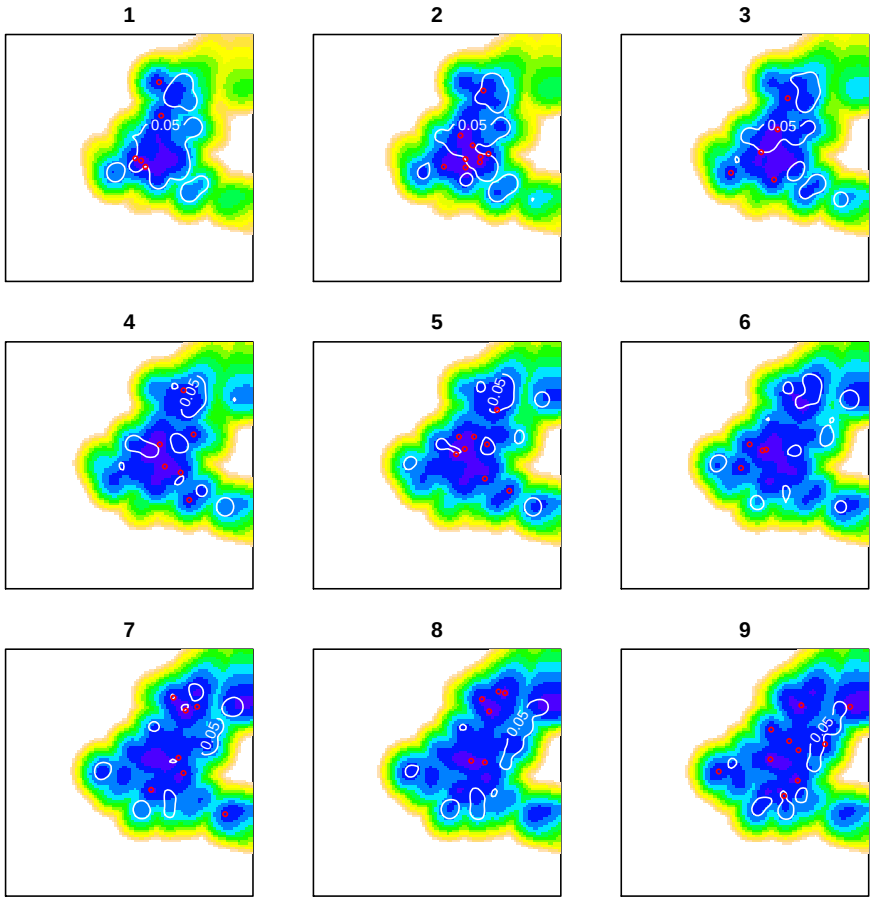

Figure 5.10: Derivative density estimation for days 1-9. LSCV bandwidth is used.

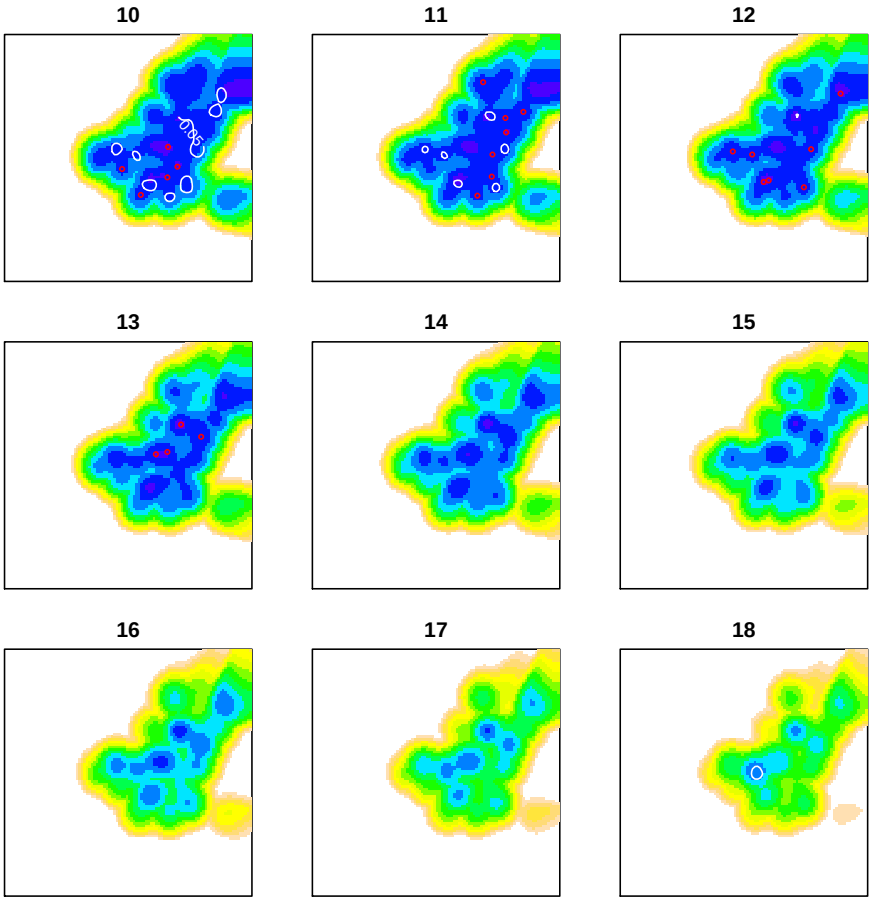

Figure 5.11: Derivative density estimation for days 10-18. LSCV bandwidth is used.

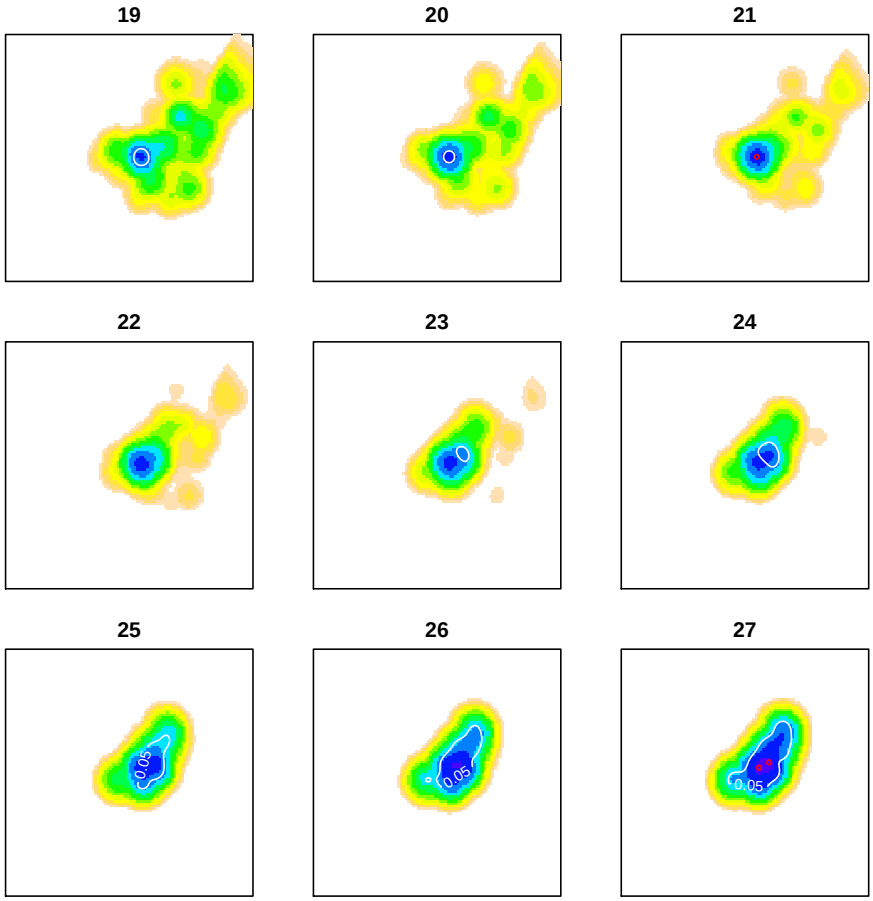

Figure 5.12: Derivative density estimation for days 19-27. LSCV bandwidth is used.

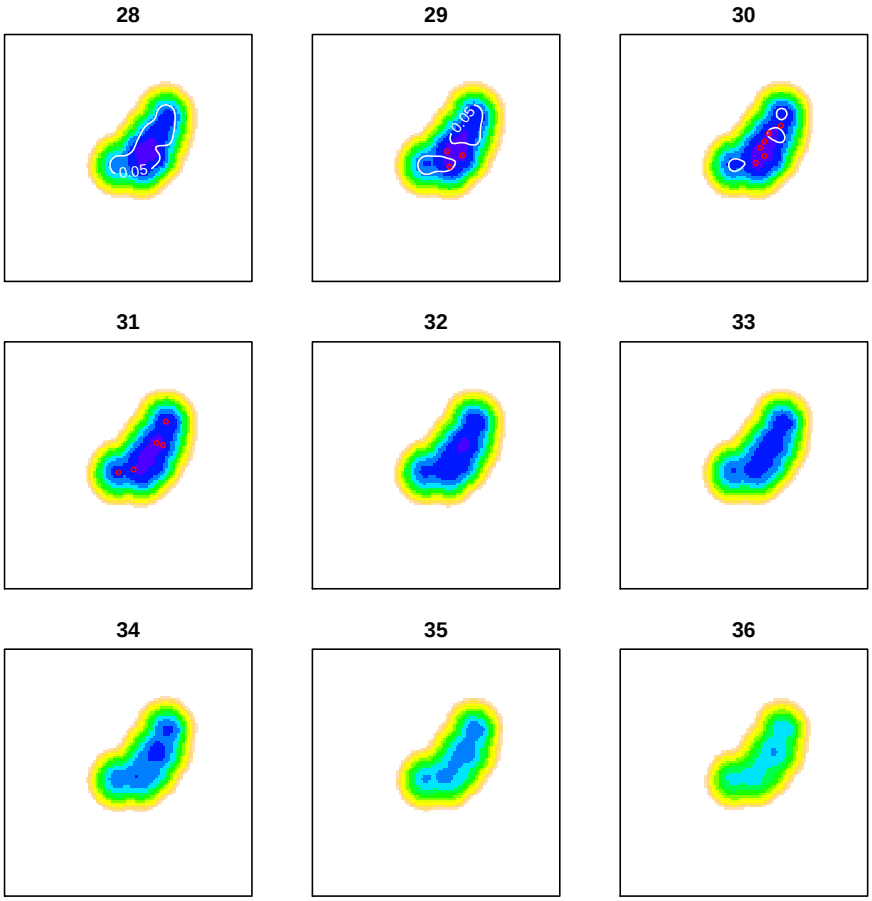

Figure 5.13: Derivative density estimation for days 28-36. LSCV bandwidth is used.

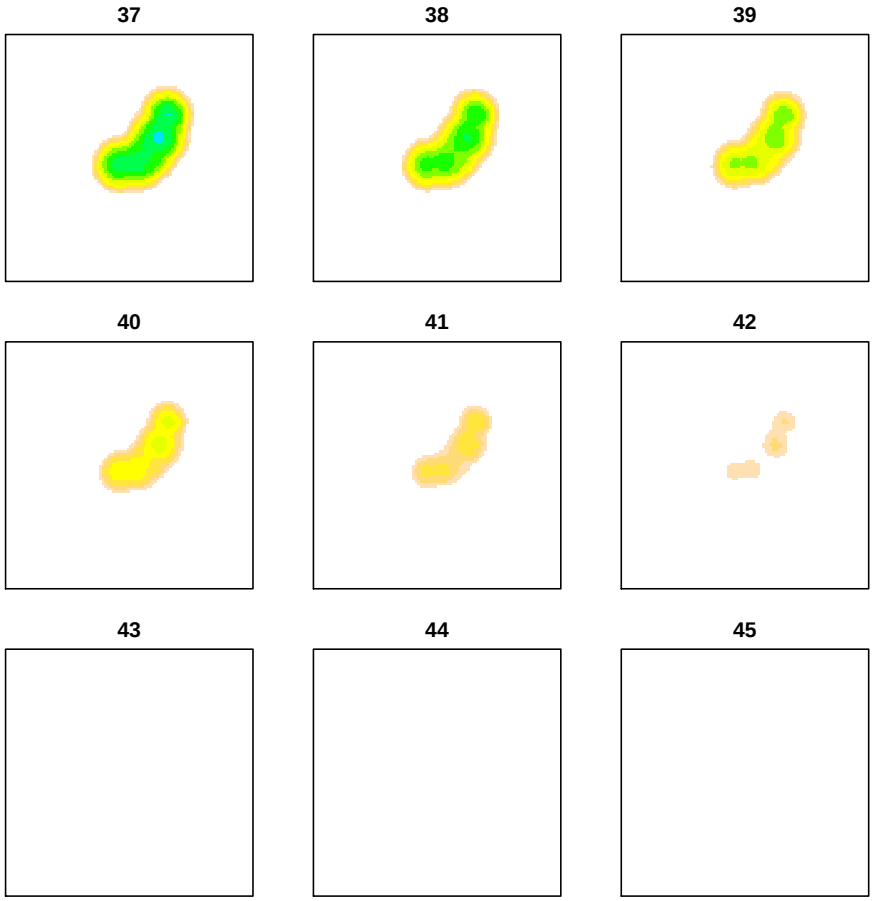

Figure 5.14: Derivative density estimation for days 37-45. LSCV bandwidth is used.

Chapter 6

Non-parametric estimation of relative risk with covariates

6.1 Introduction

Suppose in a particular epidemiological data set we have some information on a specific explanatory variable which is non-constant through the whole study. One may suspect that this particular variable might influence greatly the disease risk. We want to visualize how spatial patterns of risk vary with the covariate(s) to hand. So in this Chapter, as a novel work, we propose the generalized relative risk function,

$$r(\mathbf{x}, \mathbf{z}) = \frac{f(\mathbf{x}, \mathbf{z})}{g(\mathbf{x}, \mathbf{z})};\ \mathbf{x} \in \mathcal{R},\ \mathbf{z} \in \mathbb{R}^p. \tag{6.1.1}$$

In this chapter we use $\mathbf{z} = (z_1, ..., z_p)^T$ to represent the covariate and $\mathbf{x} = (x_1, x_2)^T$ for spatial locations. As in previous chapters, f and g represent the case and control densities respectively.

For example, in a veterinary epidemiological data set (say FMD), we might have information on farm size. The natural question then arises is how does the spatial pattern of disease change with the covariate (e.g. farm size). To answer this question we can look at the function $r(\mathbf{x}, \mathbf{z})$. According to the above definition 6.1.1, both case and control densities f and g depend on the covariate ($\mathbf{z}$). So larger farms may be concentrated in one part of the geographical region.

Note that the average of relative risk (6.1.1), is such that $r(\mathbf{x}, \mathbf{z})$ is normalized over $\mathbf{x}$ and $\mathbf{z}$, i.e., has a unit expectation with respect to control distribution of $\mathbf{x}$ and $\mathbf{z}$. That is,

$$\begin{aligned} E_{(\mathbf{x},\mathbf{z})\sim g}[r(\mathbf{x}, \mathbf{z})] &= \int_R \int_{\mathbb{R}^p} r(\mathbf{x}, \mathbf{z}) g(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &= \int_R \int_{\mathbb{R}^p} \frac{f(\mathbf{x}, \mathbf{z})}{g(\mathbf{x}, \mathbf{z})} g(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &= 1. \end{aligned}$$

An alternative is to condition on $\mathbf{z}$, so that the relative risk is given by

$$r(\mathbf{x}|\mathbf{z}) = \frac{f(\mathbf{x}|\mathbf{z})}{g(\mathbf{x}|\mathbf{z})};\ \mathbf{x} \in \mathcal{R},\ \mathbf{z} \in \mathbb{R}^p. \tag{6.1.2}$$

This is normalized with respect to $g(\mathbf{x}|\mathbf{z})$. Of course, in terms of spatial risk, (6.1.2) is just a rescaling of (6.1.1) by a factor of $g(\mathbf{z})/f(\mathbf{z})$. That is,

$$\begin{aligned} r(\mathbf{x}|\mathbf{z}) &= \frac{r(\mathbf{x}, \mathbf{z})}{r(\mathbf{z})} \\ &= r(\mathbf{x}, \mathbf{z}) \frac{g(\mathbf{z})}{f(\mathbf{z})} \end{aligned} \tag{6.1.3}$$

and

$$\begin{aligned} E_{(\mathbf{x}|\mathbf{z})\sim g}[r(\mathbf{x}|\mathbf{z})] &= \int_R \int_{\mathbb{R}^p} r(\mathbf{x}|\mathbf{z}) g(\mathbf{x}|\mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &= \int_R \int_{\mathbb{R}^p} \frac{f(\mathbf{x}|\mathbf{z})}{g(\mathbf{x}|\mathbf{z})} g(\mathbf{x}|\mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &= 1. \end{aligned}$$

The distinction between 6.1.1 and 6.1.2 is hence an issue of what we are comparing risk against. If we examine spatial plots of $r(\mathbf{x}, \mathbf{z})$ for different values of $\mathbf{z}$ then we can get a sense of overall variation in risk as the covariates change. On the other hand, plots of $r(\mathbf{x}|\mathbf{z})$ for different values of $\mathbf{z}$ will be directly comparable in terms of spatial variation of risk.

The rest of the Chapter is organized as follows. We discuss the estimation of relative risk function with covariates in Section 6.2 and then describe the kernel density estimation method in Section 6.3. We develop asymptotic results for both approaches in Section 6.4. Bandwidth selectors are briefly formulated in Section 6.5. In Section 6.6, the estimators are illustrated with FMD data of 1967 outbreak.

6.2 Relative risk function with a covariate

Let $(\mathbf{x}_1^T, \mathbf{z}_1^T)^T, \ldots, (\mathbf{x}_{n_1}^T, \mathbf{z}_{n_1}^T)^T$ denote the case data with joint density f and $(\mathbf{x}_{n_1+1}^T, \mathbf{z}_{n_1+1}^T)^T$ $, \ldots, (\mathbf{x}_n^T, \mathbf{z}_n^T)^T$ the control data with density g, where $n = n_1 + n_2$. The relative risk function, $r(\mathbf{x}, \mathbf{z})$ can be estimated by $\hat{r}$ as follows:

$$\hat{r}(\mathbf{x}, \mathbf{z}) = \frac{\hat{f}(\mathbf{x}, \mathbf{z})}{\hat{g}(\mathbf{x}, \mathbf{z})}; \ \mathbf{x} \in \mathcal{R}, \ \mathbf{z} \in \mathbb{R}^p, \tag{6.2.1}$$

which is a ratio of kernel density estimates ($\hat{f}$) and ($\hat{g}$) constructed from case and control data respectively. We discuss the form of these estimates in the next section.

Similarly,

$$\hat{r}(\mathbf{x}|\mathbf{z}) = \frac{\hat{f}(\mathbf{x}|\mathbf{z})}{\hat{g}(\mathbf{x}|\mathbf{z})};\ \mathbf{x} \in \mathcal{R},\ \mathbf{z} \in \mathbb{R}^p, \tag{6.2.2}$$

where

$$\hat{f}(\mathbf{x}|\mathbf{z}) = \frac{\hat{f}(\mathbf{x}, \mathbf{z})}{\hat{f}(\mathbf{z})} \tag{6.2.3}$$

and

$$\hat{g}(\mathbf{x}|\mathbf{z}) = \frac{\hat{g}(\mathbf{x}, \mathbf{z})}{\hat{g}(\mathbf{z})}. \tag{6.2.4}$$

The functions in 6.2.3 and 6.2.4 are defined for $\mathbf{z}$ values for which $\hat{f}(\mathbf{z}) > 0$ and $\hat{g}(\mathbf{z}) > 0$ respectively. By substituting 6.2.3 and 6.2.4 in 6.2.2, we get

$$\hat{r}(\mathbf{x}|\mathbf{z}) = \frac{\hat{r}(\mathbf{x}, \mathbf{z})}{\hat{r}(\mathbf{z})}. \tag{6.2.5}$$

This equation is equal to the eq. 6.2.1 divide by $\hat{r}(\mathbf{z})$, which is the $\int \hat{r}(\mathbf{x}, \mathbf{z}) d\mathbf{x}$.

As before, we prefer to work with log relative risk functions.

$$\begin{aligned} \hat{\rho}(\mathbf{x}, \mathbf{z}) &= \log[\hat{r}(\mathbf{x}, \mathbf{z})] \\ &= \log\left[\frac{\hat{f}(\mathbf{x}, \mathbf{z})}{\hat{g}(\mathbf{x}, \mathbf{z})}\right];\ \mathbf{x} \in \mathcal{R},\ \mathbf{z} \in \mathbb{R}^p \end{aligned}$$

and

$$\begin{aligned} \hat{\rho}(\mathbf{x}|\mathbf{z}) &= \log[\hat{r}(\mathbf{x}|\mathbf{z})] \\ &= \log\left[\frac{\hat{f}(\mathbf{x}|\mathbf{z})}{\hat{g}(\mathbf{x}|\mathbf{z})}\right];\ \mathbf{x} \in \mathcal{R},\ \mathbf{z} \in \mathbb{R}^p. \end{aligned}$$

6.3 Kernel density estimation with a covariate

In most general form, higher dimensional kernel density estimator can be expressed as

$$\hat{f}(\mathbf{x};\mathbf{z}) = n_1^{-1} \sum_{i=1}^{n_1} K_h\left(\mathbf{x}-\mathbf{x}_i\right) L_\lambda\left(\mathbf{z}-\mathbf{z}_i\right)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2})^T$, $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{ip})^T$, n_1 is the sample size, h and λ are the bandwidths based on space ($\mathbf{x}$) and the covariate ($\mathbf{z}$) respectively. $K_h(\mathbf{v}) = h^{-2}K(h^{-1}\mathbf{v})$ and $L_\lambda(\mathbf{u}) = \lambda^{-p}L(\lambda^{-1}\mathbf{u})$. K is a bivariate kernel function satisfying $\int\int K(\mathbf{x})d\mathbf{x} = 1$ and L is a p-variate kernel function satisfying $\int ... \int_p L(\mathbf{z})d\mathbf{z} = 1$. Similarly,

$$\hat{g}(\mathbf{x};\mathbf{z}) = n_2^{-1} \sum_{j=n_1+1}^{n_1+n_2} K_h\left(\mathbf{x}-\mathbf{x}_j\right) L_\lambda\left(\mathbf{z}-\mathbf{z}_j\right).$$

Therefore, from eq. (6.2.1),

$$\hat{r}(\mathbf{x},\mathbf{z}) = \frac{n_2 \sum_{i=1}^{n_1} K_h\left(\mathbf{x}-\mathbf{x}_i\right) L_\lambda\left(\mathbf{z}-\mathbf{z}_i\right)}{n_1 \sum_{j=n_1+1}^{n_1+n_2} K_h\left(\mathbf{x}-\mathbf{x}_j\right) L_\lambda\left(\mathbf{z}-\mathbf{z}_j\right)}. \tag{6.3.1}$$

6.4 Asymptotic properties

6.4.1 Bias and variance of $\hat{\rho}(\mathbf{x},\mathbf{z})$

We derive the asymptotic properties of $\hat{\rho}$ at the spatial location $\mathbf{x}$ and covariate $\mathbf{z}$ as follows. Using standard results from multivariate density estimation, we can show that

$$\text{Bias}[\hat{f}(\mathbf{x},\mathbf{z})] = \frac{1}{2}h^2\mu_2(K)\nabla^2_{\mathbf{x}}f(\mathbf{x},\mathbf{z}) + \frac{1}{2}\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}f(\mathbf{x},\mathbf{z}) + o(h^2+\lambda^2). \tag{6.4.1}$$

$$\text{Var}[\hat{f}(\mathbf{x},\mathbf{z})] = \frac{f(\mathbf{x},\mathbf{z})R(K)R(L)}{n_1h^2\lambda^p} + o\left(\frac{1}{n_1h^2\lambda^p}\right).$$

Standard manipulations then give

$$\begin{aligned}\text{Bias}[\hat{\rho}(\mathbf{x},\mathbf{z})] &= \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}f(\mathbf{x},\mathbf{z})+\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{x},\mathbf{z})}\right] \\ &\quad -\frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}g(\mathbf{x},\mathbf{z})+\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{x},\mathbf{z})}\right]+o(h^2+\lambda^2).\end{aligned}$$

and

$$\begin{aligned}\text{Var}[\hat{\rho}(\mathbf{x},\mathbf{z})] &= \frac{R(K)R(L)f(\mathbf{x},\mathbf{z})^{-1}}{n_1h^2\lambda^p}+\frac{R(K)R(L)g(\mathbf{x},\mathbf{z})^{-1}}{n_2h^2\lambda^p}+o\left(\frac{1}{n_1h^2\lambda^p}+\frac{1}{n_2h^2\lambda^p}\right) \\ &= R(K)R(L)\left[\frac{f(\mathbf{x},\mathbf{z})^{-1}}{n_1h^2\lambda^p}+\frac{g(\mathbf{x},\mathbf{z})^{-1}}{n_2h^2\lambda^p}\right]+o\left(\frac{1}{n_1h^2\lambda^p}+\frac{1}{n_2h^2\lambda^p}\right)\end{aligned}$$

As usual, we can construct tolerance contours for the $\hat{\rho}$. We consider the null hypothesis

$$H_0^1 : \rho(\mathbf{x},\mathbf{z}) = 0 \text{ (or } r(\mathbf{x},\mathbf{z}) = 1)$$

against the alternative hypothesis

$$H_1^1 : \rho(\mathbf{x},\mathbf{z}) > 0 \text{ (or } r(\mathbf{x},\mathbf{z}) > 1)$$

or one can examine $\rho(\mathbf{x},\mathbf{z}) < 0$. This creates a p-value surface over the region $\mathcal{R}$ and support of the covariate $\mathbf{z}$. As in Chapters 2, 4 and 5, we produce 5% tolerance contours using the same methodology.

An appropriate test statistic at the location $\mathbf{x}$ and covariate $\mathbf{z}$, is given by

$$Z(\mathbf{x},\mathbf{z}) = \frac{\hat{\rho}(\mathbf{x},\mathbf{z})}{\sqrt{\text{Var}[\hat{\rho}(\mathbf{x},\mathbf{z})]}}.$$

When H_0^1 is true,

$$Z(\mathbf{x},\mathbf{z}) = \frac{\hat{\rho}(\mathbf{x},\mathbf{z})}{\sqrt{\frac{R(K)R(L)g(\mathbf{x},\mathbf{z})^{-1}}{h^2\lambda^p}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}.$$

One can use pooled estimate $\hat{g}_p(\mathbf{x},\mathbf{z})$ here to replace the unknown g in the denominator of the test statistic. As usual, p-values for $Z(\mathbf{x},\mathbf{z})$ can be computed with respect to an asymptotically normal null distribution.

6.4.2 Bias and variance of $\hat{\rho}(\mathbf{x}|\mathbf{z})$

We obtain the asymptotic properties of $\hat{f}(\mathbf{x},\mathbf{z})$ and $\hat{\rho}(\mathbf{x},\mathbf{z})$ for $\mathbf{x}$ being interior to the study region $\mathcal{R}$.

Lemma 6.4.1. *Let $\boldsymbol{x}$ be an interior spatial location of the region $\mathcal{R}$ and $\boldsymbol{z}$ is a spatially varying covariate, then*

$$\begin{aligned} Bias[\hat{f}(\boldsymbol{x}|\boldsymbol{z})] &= \frac{1}{2}\left[h^2\mu_2(K)\frac{\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x},\boldsymbol{z})}{f(\boldsymbol{z})} + \lambda^2\mu_2(L)\frac{\nabla_{\boldsymbol{z}}^2 f(\boldsymbol{x},\boldsymbol{z})}{f(\boldsymbol{z})} - \lambda^2\mu_2(L)\frac{f(\boldsymbol{x},\boldsymbol{z})\nabla_{\boldsymbol{z}}^2 f(\boldsymbol{z})}{f(\boldsymbol{z})^2}\right] \\ &\quad + o\left(h^2+\lambda^2\right). \end{aligned}$$

and

$$Var[\hat{f}(\boldsymbol{x}|\boldsymbol{z})] = \frac{f(\boldsymbol{x}|\boldsymbol{z})R(K)R(L)}{f(\boldsymbol{z})n_1h^2\lambda^p} + o\left(\frac{1}{n_1h^2\lambda^p}\right).$$

Theorem 6.4.2. *If $\boldsymbol{x}$ be an interior point of $\mathcal{R}$ and $\boldsymbol{z}$ is a spatially varying covariate then*

$$
\begin{aligned}
Bias[\hat{\rho}(\boldsymbol{x}|\boldsymbol{z})] &= \frac{1}{2}h^2\mu_2(K)\left[\frac{\nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x},\boldsymbol{z})}{f(\boldsymbol{x},\boldsymbol{z})} - \frac{\nabla_{\boldsymbol{x}}^2 g(\boldsymbol{x},\boldsymbol{z})}{g(\boldsymbol{x},\boldsymbol{z})}\right] \\
&\quad + \frac{1}{2}h^2\mu_2(L)\left[\frac{\nabla_{\boldsymbol{z}}^2 f(\boldsymbol{x},\boldsymbol{z})}{f(\boldsymbol{x},\boldsymbol{z})} - \frac{\nabla_{\boldsymbol{z}}^2 g(\boldsymbol{x},\boldsymbol{z})}{g(\boldsymbol{x},\boldsymbol{z})}\right] \\
&\quad - \frac{1}{2}\lambda^2\mu_2(L)\left[\frac{\nabla_{\boldsymbol{z}}^2 f(\boldsymbol{z})}{f(\boldsymbol{z})} - \frac{\nabla_{\boldsymbol{z}}^2 g(\boldsymbol{z})}{g(\boldsymbol{z})}\right] + o(h^2+\lambda^2)
\end{aligned}
$$

and

$$
Var[\hat{\rho}(\boldsymbol{x}|\boldsymbol{z})] = \frac{R(K)R(L)}{h^2\lambda^p}\left[\frac{1}{n_1 f(\boldsymbol{x},\boldsymbol{z})} + \frac{1}{n_2 g(\boldsymbol{x},\boldsymbol{z})}\right] + o(n_1^{-1}h^{-2}\lambda^{-p} + n_2^{-1}h^{-2}\lambda^{-p}).
$$

One noteworthy result is that the variance of $\hat{\rho}(\mathbf{x}|\mathbf{z})$ is similar to that of $\hat{\rho}(\mathbf{x},\mathbf{z})$. Proofs of the Lemma 6.4.1 and the Theorem 6.4.2 are given in the Appendix D.1 and D.2 respectively. If $\mathbf{x}$ is close to the boundary of the region then these estimates need edge correction as discussed in the previous Chapters.

We construct the tolerance contours for $\rho(\mathbf{x}|\mathbf{z})$ over the region $\mathcal{R}$ when the covariate $\mathbf{z}$ is given by considering the null hypothesis

$$
H_0^2 : \rho(\mathbf{x}|\mathbf{z}) = 0 \text{ (or } r(\mathbf{x}|\mathbf{z}) = 1)
$$

against the alternative hypothesis

$$
H_1^2 : \rho(\mathbf{x}|\mathbf{z}) > 0 \text{ (or } r(\mathbf{x}|\mathbf{zz}) > 1).
$$

An appropriate test statistic at the location $\mathbf{x}$ when the covariate is $\mathbf{z}$,

$$
Z(\mathbf{x},\mathbf{z}) = \frac{\hat{\rho}(\mathbf{x}|\mathbf{z})}{\sqrt{\mathrm{Var}[\hat{\rho}(\mathbf{x}|\mathbf{z})]}}.
$$

When H_0^2 is true,

$$Z(\mathbf{x},\mathbf{z}) = \frac{\hat{\rho}(\mathbf{x}|\mathbf{z})}{\sqrt{\frac{R(K)R(L)}{h^2\lambda^p}\left[\frac{1}{n_1 f(\mathbf{x},\mathbf{z})} + \frac{1}{n_2 g(\mathbf{x},\mathbf{z})}\right]}}.$$

Here we cannot assume that $f(\mathbf{x},\mathbf{z}) = g(\mathbf{x},\mathbf{z})$ and so we use pilot estimates for f and g, which can be computed from multivariate kernel density estimates. Also it is necessary to apply edge correction criteria for the estimates when the observations are close to the boundary.

6.5 Bandwidth Selection

Using standard properties of bias and variance, we derive from eq. (5.7.1),

$$MISE(\hat{\rho}(\mathbf{x};\mathbf{z})) = \int\int \left[\text{Bias}\{\widehat{\rho}(\mathbf{x},\mathbf{z})\}\right]^2 d\mathbf{x}d\mathbf{z} + \int\int \text{Var}\{\widehat{\rho}(\mathbf{x},\mathbf{z})\}d\mathbf{x}d\mathbf{z}.$$

By minimizing MISE, we get the optimal bandwidth combination,

$$(h,\lambda)_{\text{opt}} = \arg\min_{h,\lambda}[MISE(\hat{\rho}(\mathbf{x},\mathbf{z}))].$$

Subjective bandwidth and automatic bandwidth can be applied here. We prepare using the LSCV bandwidth selector since it performs relatively well in density ratio method. We derive the LSCV bandwidth formulae from eq. (5.7.3) as follows:

$$\begin{aligned}\widehat{CV}(h,\lambda) = &-\int_{\mathcal{R}}\int_{\mathbb{R}^p}\{\hat{\rho}_{h,\lambda}(\mathbf{x},\mathbf{z})\}^2 d\mathbf{x}d\mathbf{z} - 2n_1^{-1}\sum_{i=1}^{n_1}\log\left[\frac{\hat{f}_{h,\lambda}^{-i}(\mathbf{x}_i,\mathbf{z}_i)}{\hat{g}_{h,\lambda}(\mathbf{x}_i,\mathbf{z}_i)}\right]\{\hat{f}^{-i}(\mathbf{x}_i,\mathbf{z}_i)\}^{-1} \\ &+ 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2}\log\left[\frac{\hat{f}_{h,\lambda}(\mathbf{x}_j,\mathbf{z}_j)}{\hat{g}_{h,\lambda}^{-j}(\mathbf{x}_j,\mathbf{z}_j)}\right]\{\hat{g}^{-j}(\mathbf{x}_j,\mathbf{z}_j)\}^{-1}.\end{aligned}$$

The optimal bandwidth combination (h, λ) can be obtained by minimizing the objective function $\widehat{CV}$ over h and λ.

Now we look at how this bandwidth selector works for $\hat{\rho}(\mathbf{x}|\mathbf{z})$. The MISE is defined as follows:

$$MISE(\hat{\rho}(\mathbf{x}|\mathbf{z})) = \int\int \left[\text{Bias}\{\widehat{\rho}(\mathbf{x}|\mathbf{z})\}\right]^2 d\mathbf{x}d\mathbf{z} + \int\int \text{Var}\{\widehat{\rho}(\mathbf{x}|\mathbf{z})\}d\mathbf{x}d\mathbf{z}.$$

then we get the optimal bandwidth

$$(h, \lambda)_{\text{opt}} = \arg\min_{h,\lambda} MISE[\hat{\rho}(\mathbf{x}|\mathbf{z})].$$

The LSCV bandwidth selector can be defined as the minimizer of

$$\begin{aligned}\widehat{CV}(h, \lambda) = &- \int\int \{\hat{\rho}_{h,\lambda}(\mathbf{x}|\mathbf{z})\}^2 d\mathbf{x}d\mathbf{z} - 2n_1^{-1}\sum_{i=1}^{n_1} \log\left[\frac{\hat{f}^{-i}_{h,\lambda}(\mathbf{x}_i|\mathbf{z}_i)}{\hat{g}_{h,\lambda}(\mathbf{x}_i|\mathbf{z}_i)}\right] \{\hat{f}^{-i}(\mathbf{x}_i|\mathbf{z}_i)\}^{-1} \\ &+ 2n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} log\left[\frac{\hat{f}_{h,\lambda}(\mathbf{x}_j|\mathbf{z}_j)}{\hat{g}^{-j}_{h,\lambda}(\mathbf{x}_j|\mathbf{z}_j)}\right] \{\hat{g}^{-j}(\mathbf{x}_j|\mathbf{z}_j)\}^{-1}.\end{aligned}$$

The optimal bandwidth combination (h, λ) can be obtained by minimizing the objective function $\widehat{CV}$ over h and λ.

6.6 Real application: The 2001 outbreak of FMD

These particular counts of cases of FMD were available for parishes within the county of Cumbria, Northern England for the period February 2001 to August 2001, as analyzed by Lawson and Zhou (2005). The dataset includes locations of individual farms (x.coord,y.coord) which is either a case farm or a control farm (status, 1 for case farm, 0 for control farm), their date of infection (days, if a case otherwise coded to 241)

and the total population (total), which is the total number of cattles, sheep, goats, pigs and mixed animals. We use this total animal population (total), z as a covariate for the relative risk function, r.

The data include the locations of 418 cases of FMD identified within the study region during February 2001 to August 2001, along with the locations of 2395 controls. The spatial distribution is shown as separate plots for cases (left panel) and controls (right panel) in Figure 6.1. It is clear that case farms are clustered in the northwest part of the region.

Figure 6.2 shows the log relative risk estimators ($\hat{\rho}(\mathbf{x}|\mathbf{z})$) for different farm sizes using LSCV bandwidth, in which spatial and temporal components, $h = 27512$ and $\lambda = 118$ respectively, are used. Solid lines represent the 5% tolerance contours. This plot produces sensible tolerance contours in the northwest region. In particular, the pattern of spatial risk seems to change with farm size.

6.7 Conclusion

The use of relative risk estimation (using density ratio method) in geographical epidemiology including a spatially varying covariate is a potentially useful tool in epidemiology.

In this chapter, we have proposed two versions of the log relative risk function with a covariate. One can use either depending on the comparison that one wishes to focus on. We propose estimators and obtain the asymptotic properties of bias and variance

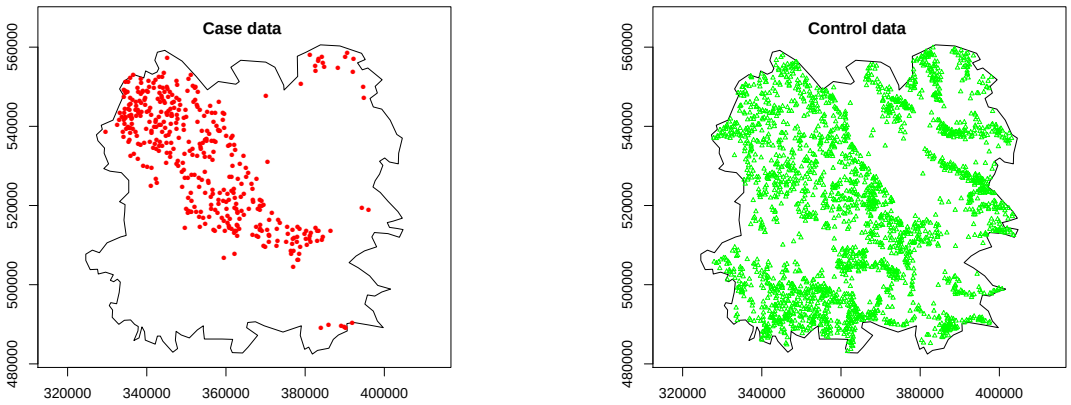

Figure 6.1: Cases and controls for the FMD dataset, including the defined region. Each bullet point represents a farm.

of $\hat{\rho}$ in both cases.

The selection of smoothing parameter is critical in all kernel smoothing estimation problems. We obtain a formulae using LSCV bandwidth selector in density ratio estimation. As FMD data set is used to illustrate log relative risk estimates with tolerance contours to identify the regions having significantly elevated risk where we use LSCV as the bandwidth.

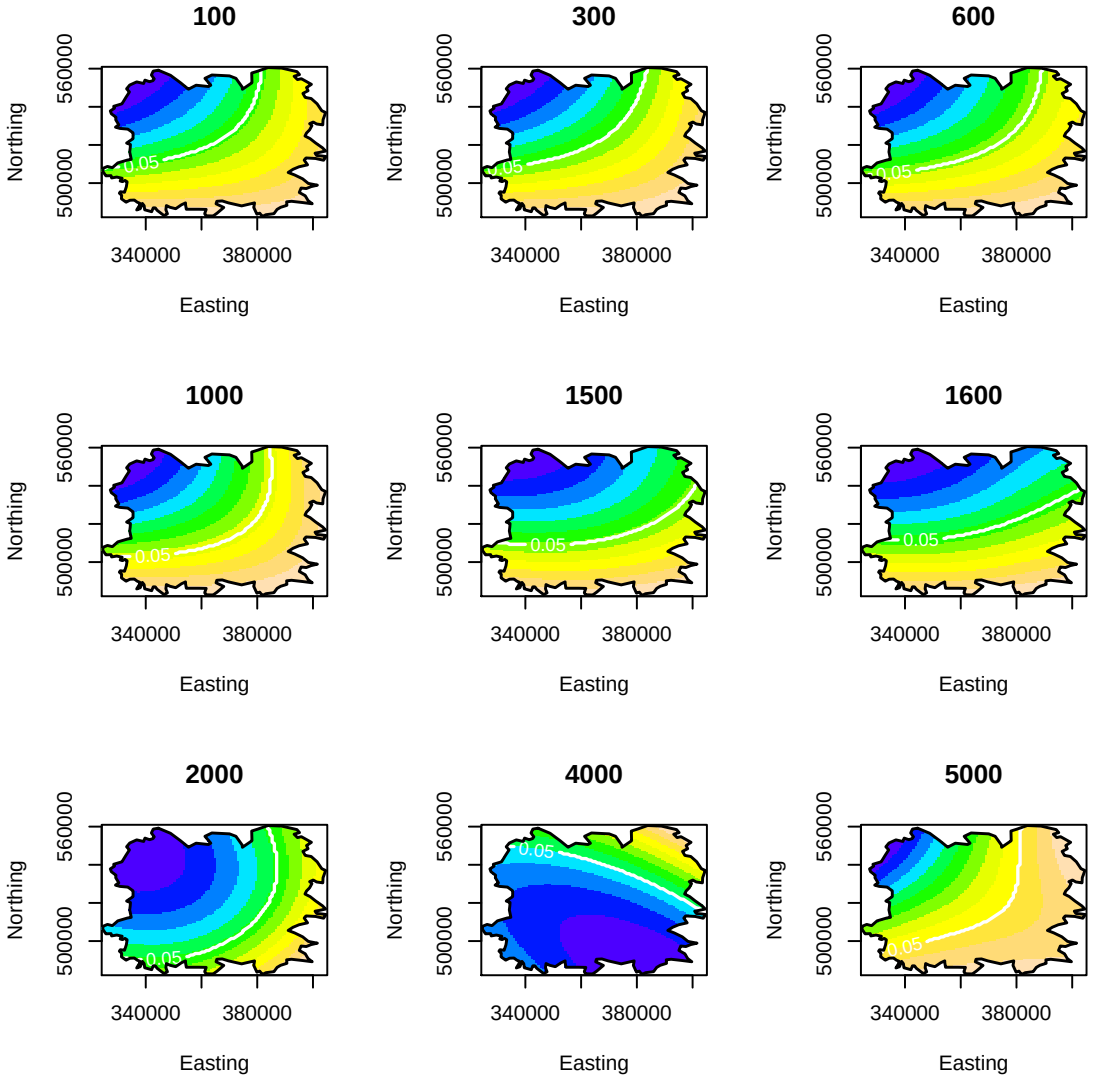

Figure 6.2: Heatplots of FMD relative risk surfaces (on log scale), with 5% tolerance contours (solid internal lines). The covariate (total population) is displayed as the title at each plot. LSCV bandwidth is used.

Chapter 7

General Discussion

In this Chapter, I summarize my research work in Section 7.1 and describe possible future work in Section 7.2.

7.1 Summary of my work

This PhD thesis is primarily based on the work on non-parametric estimation of geographical relative risk function ($r(\mathbf{x}) = \frac{f(\mathbf{x})}{g(\mathbf{x})}; \mathbf{x} \in \mathcal{R}$), which is a useful tool for investigating the spatial variation of disease risk based on case and control data. The standard method of estimating this function in a geographical region is known as density ratio which is defined as the ratio of bivariate kernel density estimates (Bithel, 1990), constructed from the locations of cases and the controls respectively. Following the argument made by Kelsall and Diggle (1995 a,b), we focused on estimating the relative risk function on log scale ($\rho(\mathbf{x}) = \log r(\mathbf{x}); \mathbf{x} \in \mathcal{R}$).

The intention of this thesis is twofold. First, we work on spatially varying relative

risk surfaces. Then we extend it to time varying spatial relative risk function and also investigate the properties of relative risk function when there is a spatially varying covariate.

The kernel density estimator ($\hat{f}$) for a generic function f depends on the bandwidth (h) and the kernel (K). There is no much difference in performance between kernels, but the choice of bandwidth is crucial. We conducted a simulation study to compare cross-validation bandwidth selectors LCV and LSCV in the estimation of ρ using density ratio method. We found the LSCV method to be preferable, although its performance was moderately sensitive to the choice of search interval for the optimal bandwidth.

The density ratio estimator of the relative risk function has proved adept at detecting approximately circular or elliptical areas of having high risk often around chosen point sources of risk (Hazelton and Davies, 2009). However, this method can be less natural when the relative risk has a more global trend such as a linear source of risk. So we examined local linear regression as an alternative to estimate the log-relative risk function.

We carried out a simulation study to compare the performance of the density ratio estimator with that of the local linear estimator in the estimation of relative risk. Here we chose the synthetic problems, generated by a 2^4 factorial design. Overall, our results suggests that the density ratio is far better in most of the scenarios in the estimation of relative risk function while local linear method performs well if there is

a linear source of risk.

Bandwidth selection is crucial in every kernel smoothing problem. Our overall recommendation about bandwidth selectors is that LSCV is the best choice for density ratio estimation while plug-in bandwidth selector is preferred with local linear estimator where we expect long range trends in risks.

We extend the previous work to the context of space-time. According to previous research, fields such as epidemiology, geology, meteorology and ecology generate data which consist of both spatial location and time components, but the change in risk pattern in time dimension has often been neglected in the research. So nowadays there is a growing interest in integrating temporal information in the relevant research. Therefore, we define spatio-temporal relative risk function over a geographical region. After that multivariate kernel density estimator was introduced in the context of space-time to compute unknown densities. We obtained the asymptotic properties of the kernel density estimators and relative risk estimators. Also, we describe the tolerance contours of relative risk estimators. It is generally accepted that the choice of bandwidth is critical in any dimensional kernel density estimators. So here we discuss cross-validation bandwidth selectors, LCV and LSCV to compute optimal spatial and temporal bandwidths.

We expand the previous work into time derivative density estimation and its properties since it is important to distinguish the rate of change of relative risk of a disease over time. We derived formulas for the asymptotic properties of these time derivative

estimators.

Our purpose in proposing the spatio-temporal relative risk function is primarily to provide a tool for visualizing the evaluation of a disease outbreak in time and space. We illustrate this using FMD data, from the 1967 outbreak in Cumbria, UK. We produce log relative risk estimators and its time derivative estimators and use 5% tolerance contours to highlight significant features of the elevated risk surfaces.

The geographical relative risk function including a spatially varying covariate has been not studied in the literature. We propose such a function so as to visualize how the patterns of risk vary with covariates. We defined kernel estimates for this function, and obtained its asymptotic properties. We discuss $LSCV$ to estimate the bandwidth. We apply this methodology to data from the 2001 outbreak of FMD in the UK by using farm population as a covariate.

7.2 Suggestions for future work

The dissertation focuses on the non-parametric estimation. But in some situations, semi-parametric statistical models might be appropriate. For example, we might have a parametric model for the effect of some point source of risk, or some covariate, that we want to incorporate into or otherwise non-parametric approach to estimation.

We have developed a plug-in bandwidth selector in the estimation of local linear method. The possibility of using plug-in bandwidth in density ratio selector is attractive, though a challenging research problem.

Selection of bandwidths for spatio-temporal relative risk estimation and the estimation with covariates is another research direction that warrants attention. Again the development of a plug-in bandwidth selection methodology is attractive in principles, although challenging. We work on density ratio methodology in spatio-temporal relative risk estimation, but wish to investigate other possible methods starting from local linear method. We might also consider semi-parametric approaches in which at least some of the temporal variation is modelled parametrically. For example, we might introduce parametric models to handle seasonality.

Appendix

A.1: Proof of Theorem 2.4.1

Following Kelsall & Diggle (1995a), we wish to estimate $\rho_{h_1,h_2}(x) = \log\{f_{h_1}(x)/g_{h_2}(x)\}$, where n_1, n_2 are sample sizes of cases and controls while h_1, h_2 are respective bandwidths. For our asymptotic approximations to be valid, we assume that n_1 and n_2 are large and that h_1 and h_2 are small, decreasing as the sample sizes increase. We also assume that f and g are both bounded away from zero and twice differentiable on the interval I.

We define the relative error term for $\hat{f}_{h_1}(x)$ as

$$\epsilon_f(x; h_1) = \{\hat{f}_{h_1}(x) - f(x)\}/f(x),$$

for $x \,\epsilon\, I$, and similarly $\epsilon_g(x; h_2)$ for $\hat{g}_{h_2}(x)$. According to the assumptions we made, these error terms will be small. Rearranging the above error term gives

$$\hat{f}_{h_1}(x) = f(x)\{1 + \epsilon_f(x; h_1)\},$$

and similarly for $\hat{g}_{h_2}(x)$. Using first order Taylor expansion of $\hat{\rho}_{h_1,h_2}(x)$ function, we

get

$$\begin{aligned}\hat{\rho}_{h_1,h_2}(x) &= \log\left[\frac{f(x)}{g(x)}\frac{(1+\epsilon_f(x;h_1))}{(1+\epsilon_g(x;h_2))}\right] \\ &= \rho(x)+\log(1+\epsilon_f(x;h_1))-\log(1+\epsilon_g(x;h_2)) \\ &= \rho(x)+\epsilon_f(x;h_1)-\epsilon_g(x;h_2)+o(\epsilon^2),\end{aligned}$$

where ϵ denotes either $\epsilon_f(x;h_1)$ or $\epsilon_g(x;h_2)$. Using the asymptotic formulae in equation (2.4.1) of $\hat{f}$ and similarly for $\hat{g}$ we get

$$E[\hat{\rho}_{h_1,h_2}(x)] = \rho(x)+\mu_2(K)\frac{1}{2}\left[h_1^2\frac{f^{"}(x)}{f(x)}-h_2^2\frac{g^{"}(x)}{g(x)}\right]+o\left(h_1^2+h_2^2\right)$$

and substituting the equation (2.4.2), we get

$$\mathrm{Var}[\hat{\rho}_{h_1,h_2}(x)] = R(K)\left[n_1^{-1}h_1^{-1}f(x)^{-1}+n_2^{-1}h_2^{-1}g(x)^{-1}\right]+o\left(n_1^{-1}h_1^{-1}+n_2^{-1}h_2^{-1}\right).$$

A.2 : Proof of Theorem 2.4.6

Following Kelsall & Diggle (1995b), we can derive the bias and variance of $\hat{\rho}(\mathbf{x})$ as follows: As

$$\hat{\rho}(\mathbf{x}) = \log[\hat{f}(\mathbf{x})]-\log[\hat{g}(\mathbf{x})]$$

Let's define the relative error term, $\epsilon_f(\mathbf{x})$ of $\hat{f}(\mathbf{x})$

$$\epsilon_f(\mathbf{x}) = \frac{\hat{f}(\mathbf{x})-f(\mathbf{x})}{f(\mathbf{x})}$$

Therefore,

$$\hat{f}(\mathbf{x}) = f(\mathbf{x})\{1+\epsilon_f(\mathbf{x})\}.$$

Similarly,

$$\hat{g}(\mathbf{x}) = g(\mathbf{x})\{1+\epsilon_g(\mathbf{x})\}.$$

Therefore,

$$\begin{aligned}\hat{\rho}(\mathbf{x}) &= \log\left[\frac{f(\mathbf{x})\{1+\epsilon_f(\mathbf{x})\}}{g(\mathbf{x})\{1+\epsilon_g(\mathbf{x})\}}\right] \\ &= \rho(x) + \log\{1+\epsilon_f(\mathbf{x})\} - \log\{1+\epsilon_g(\mathbf{x})\} \\ &\approx \rho(\mathbf{x}) + \epsilon_f(\mathbf{x}) - \epsilon_g(\mathbf{x}).\end{aligned}$$

Then the expectation of $\hat{\rho}(\mathbf{x})$ is

$$E[\hat{\rho}(\mathbf{x})] \approx \rho(\mathbf{x}) + E[\epsilon_f(\mathbf{x})] - E[\epsilon_g(\mathbf{x})].$$

where

$$\begin{aligned}E[\epsilon_f(\mathbf{x})] &= \frac{E[\hat{f}(\mathbf{x}) - f(\mathbf{x})]}{f(\mathbf{x})} \\ &\approx \frac{1}{2}h^2\mu_2(K)\nabla^2 f(\mathbf{x})/f(\mathbf{x})\end{aligned}$$

Similarly,

$$E[\epsilon_g(\mathbf{x})] \approx \frac{1}{2}h^2\mu_2(K)\nabla^2 g(\mathbf{x})/g(\mathbf{x}).$$

Therefore,

$$E[\hat{\rho}(\mathbf{x})] \approx \rho(\mathbf{x}) + \frac{1}{2}h^2\mu_2(K)\left[\nabla^2 f(\mathbf{x})/f(\mathbf{x}) - \nabla^2 g(\mathbf{x})/g(\mathbf{x})\right].$$

The variance term is

$$\begin{aligned}\mathrm{Var}[\hat{\rho}(\mathbf{x})] &= \mathrm{Var}(\epsilon_f) + \mathrm{Var}(\epsilon_g) + \mathrm{Cov}(\epsilon_f, \epsilon_g) \\ &= \frac{\mathrm{Var}\{\hat{f}(\mathbf{x})\}}{f(\mathbf{x})^2} + \frac{\mathrm{Var}\{\hat{g}(\mathbf{x})\}}{g(\mathbf{x})^2} (\because \text{By assuming the error terms are independent}) \\ &\approx \frac{R(K)}{n_1 h^2 f(\mathbf{x})} + \frac{R(K)}{n_2 h^2 g(\mathbf{x})} \\ &\approx \frac{R(K)}{h^2}[n_1^{-1} f(\mathbf{x})^{-1} + n_2^{-1} g(\mathbf{x})^{-1}]\end{aligned}$$

Assume that the bandwidths of case and control densities are h_1 and h_2. Then

$$E[\hat{\rho}(\mathbf{x})] = \rho(\mathbf{x}) + \frac{1}{2}\mu_2(K)\left[h_1^2\frac{\nabla^2 f(\mathbf{x})}{f(\mathbf{x})} - h_2^2\frac{\nabla^2 g(\mathbf{x})}{g(\mathbf{x})}\right] + o(h_1^2 + h_2^2)$$

and

$$\mathrm{Var}[\hat{\rho}(\mathbf{x}, h_1, h_2)] \;=\; R(K)\left[n_1^{-1}h_1^{-2}f(\mathbf{x})^{-1} + n_2^{-1}h_2^{-2}g(\mathbf{x})^{-1}\right] + o(n_1^{-1}h_1^{-2} + n_2^{-1}h_2^{-2}).$$

B: Direct Proof of the Theorem 4.4.1

Proof. Kernel weighted log-likelihood function for local linear regression, given in the equation (4.2.3) is used to derive the asymptotic bias and variance of the local linear relative risk estimator. The equation (4.2.3) is written in matrix form as follows:

$$\bar{L}(\beta; h; \mathbf{x}) = \mathbf{y}^T W_{\mathbf{x}}(\mathbf{X}_{\mathbf{x}}\beta) - 1^T W_{\mathbf{x}} b(\mathbf{X}_{\mathbf{x}}\beta)$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\quad W_{\mathbf{x}} = \mathrm{diag}(\kappa_1, \ldots, \kappa_n)$, where $\kappa_i = K_h(\mathbf{X}_i - \mathbf{x})$. and $\beta = (\beta_0, \beta_1, \beta_2)^T$. Here $\mathbf{X}_i = (X_{1i}, X_{2i}), i = 1(.)n$ represents the locations of data while $\mathbf{x} = (x_{01}, x_{02})$, any generic point. h is the smoothing parameter. $\mathbf{X}_{\mathbf{x}}$ is known as the $n \times 3$ design matrix

$$\mathbf{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (X_{11} - x_{01}) & (X_{21} - x_{02}) \\ \vdots & \vdots & \vdots \\ 1 & (X_{1n} - x_{01}) & (X_{2n} - x_{02}) \end{pmatrix}.$$

Here $b(\mathbf{x}) = \log(1 + e^{\mathbf{x}})$ corresponds to binary regression with a logit link function. Note that the function b is to be interpreted element-wise. To obtain the local polynomial estimator of relative risk function, we need to maximize the function $\bar{L}$. By

differentiating the above $\bar{L}$ function with respect to β,

$$\begin{aligned}
\bar{L}^{'}(\beta, h, \mathbf{x}) &= \mathbf{y}^T W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}} - 1^T W_{\mathbf{x}} \text{diag}\{b^{'}(\mathbf{X}_{\mathbf{x}}\beta)\}\mathbf{X}_{\mathbf{x}} \\
&= \left[\mathbf{y}^T W_{\mathbf{x}} - (\kappa_1, \ldots, \kappa_n)\text{diag}\{b^{'}(\mathbf{X}_{\mathbf{x}}\beta)\}\right] \mathbf{X}_{\mathbf{x}} \\
&= \left[\mathbf{y}^T W_{\mathbf{x}} - b^{'}(\mathbf{X}_{\mathbf{x}}\beta)^T \text{ diag}(\kappa_1, \ldots, \kappa_n)\right] \mathbf{X}_{\mathbf{x}} \\
&= \left[\mathbf{y} - b^{'}(\mathbf{X}_{\mathbf{x}}\beta)\right]^T W_{\mathbf{x}} X_{\mathbf{x}}
\end{aligned}$$

We get the second derivative of $\bar{L}$ with respect to β.

$$\begin{aligned}
d^2\bar{L}(\beta; h, \mathbf{x}) &= d\left[\mathbf{y} - b^{'}(\mathbf{X}_{\mathbf{x}}\beta)\right]^T W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}} d\beta \\
&= -\left[\text{diag}(b^{''}(\mathbf{X}_{\mathbf{x}}\beta))\mathbf{X}_{\mathbf{x}} d\beta\right]^T W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}} d\beta \\
&= (d\beta)^T \mathbf{X}_{\mathbf{x}}^T \left[-\text{diag}(b^{''}(\mathbf{X}_{\mathbf{x}}\beta))\right] W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}} d\beta
\end{aligned}$$

By the definition of Hessian matrix,

$$H(\bar{L}(\beta; h, \mathbf{x})) = -\mathbf{X}_{\mathbf{x}}^T \text{diag}(b^{''}(\mathbf{X}_{\mathbf{x}}\beta)) W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}}.$$

That is,

$$\bar{L}^{''}(\beta; h, \mathbf{x}) = -\mathbf{X}_{\mathbf{x}}^T \text{diag}(b^{''}(\mathbf{X}_{\mathbf{x}}\beta)) W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}}$$

Multivariate Taylor's theorem gives

$$0 = \bar{L}^{'}(\hat{\beta}, h, \mathbf{x}) + \left[\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x})\right]^T \bar{L}^{''}(\hat{\beta}; h, \mathbf{x}) + \text{Remainder term}$$

$$\begin{aligned}
\Rightarrow \left[\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x})\right]^T &= -\bar{L}^{'}(\hat{\beta}, h, \mathbf{x})\{\bar{L}^{''}(\hat{\beta}; h, \mathbf{x})\}^{-1} + \text{Remainder term} \\
\Rightarrow \hat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) &= -\{\bar{L}^{''}(\hat{\beta}; h, \mathbf{x})\}^{-1} \mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} (\mathbf{y} - b^{'}(\mathbf{X}_{\mathbf{x}}\hat{\beta})) + \text{Remainder term}
\end{aligned}$$

By substituting $\bar{L}^{'}$ and $\bar{L}^{''}$, we get

$$\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) = \left[\mathbf{X}_{\mathbf{x}}^T \text{diag}(b^{''}(\mathbf{X}_{\mathbf{x}}\hat{\beta})) W_{\mathbf{x}} \mathbf{X}_{\mathbf{x}}\right]^{-1} \mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} (\mathbf{y} - b^{'}(\mathbf{X}_{\mathbf{x}}\hat{\beta}))$$

Since the estimator of $\beta(\mathbf{x})$ is the intercept coefficient, we obtain

$$E\left[\hat{\beta}(\mathbf{x})-\beta(\mathbf{x})|\mathbf{X}_1,..,\mathbf{X}_n\right]=\mathbf{e}_1^T\left[\mathbf{X}_\mathbf{x}^T\text{diag}(b''(\mathbf{X}_\mathbf{x}\hat{\beta}))W_\mathbf{x}\mathbf{X}_\mathbf{x}\right]^{-1}\mathbf{X}_\mathbf{x}^TW_\mathbf{x}E[\mathbf{y}-b'(\mathbf{X}_\mathbf{x}\hat{\beta})|\mathbf{X}_1,...,\mathbf{X}_n] \tag{7.0.1}$$

where $\mathbf{e}_1$ is the 2×1 vector having 1 in the first entry and zero elsewhere.

$\mathbf{y}-b'(\mathbf{X}_\mathbf{x}\hat{\beta})$ can be expanded as follows:

$$\mathbf{y}-b'(\mathbf{X}_\mathbf{x}\hat{\beta})=\begin{bmatrix} y_1-b'(\mathbf{X}_\mathbf{x}\hat{\beta}_1)\\ \vdots \\ y_n-b'(\mathbf{X}_\mathbf{x}\hat{\beta}_n)\end{bmatrix}$$

where

$$\begin{aligned} b'(\mathbf{X}_\mathbf{x}\beta_i) &= \frac{\exp(\mathbf{X}_\mathbf{x}\beta_i)}{1+\exp(\mathbf{X}_\mathbf{x}\beta_i)}\\ &= \frac{1}{\exp(-\mathbf{X}_\mathbf{x}\beta_i)+1}. \end{aligned} \tag{7.0.2}$$

Here $\mathbf{X}_\mathbf{x}\beta_i=\beta_0+\beta_1(x_{i1}-x_{01})+\beta_2(x_{i2}-x_{02})$ and from equation (4.4.1),

$$\log\left[\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right]=\mathbf{X}_\mathbf{x}\beta_i.$$

By applying the second order Taylor expansion, we get

$$\log\left[\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right]=\log\left[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right]+(\mathbf{x}_i-\mathbf{x})^T\mathcal{D}(\mathbf{x})+\frac{1}{2}(\mathbf{x}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{x}_i-\mathbf{x})+...$$

where $\mathcal{D}$ and $\mathcal{H}$ represent the first order partial derivatives and the Hessian matrix

respectively. By substituting this in equation (7.0.2),

$$\begin{aligned}
b^{'}(\mathbf{X}_{\mathbf{x}}\beta_i) &= \frac{1}{\exp\left[-\log[\frac{p(\mathbf{x})}{1-p(\mathbf{x})}] + \frac{1}{2}(\mathbf{x}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{x}_i-\mathbf{x}) - o\|\mathbf{x}_i-\mathbf{x}\|^2\right]+1} \\
&= \frac{p(\mathbf{x})}{[1-p(\mathbf{x})]\exp\left[\frac{1}{2}(\mathbf{x}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{x}_i-\mathbf{x}) - o\|\mathbf{x}_i-\mathbf{x}\|^2\right]+p(\mathbf{x})} \\
&= \frac{p(\mathbf{x})}{[1-p(\mathbf{x})][1+1/2\ (.) - o\|.\|^2]+p(\mathbf{x})} \\
&= \frac{p(\mathbf{x})}{1+[1-p(\mathbf{x})][1/2(.) - o\|.\|^2]} \\
&= p(\mathbf{x})\left[1-(1-p(\mathbf{x}))\left[\frac{1}{2}[\mathbf{x}_i-\mathbf{x}]^T\mathcal{H}(\mathbf{x})[\mathbf{x}_i-\mathbf{x}] - o\|\mathbf{x}_i-\mathbf{x}\|^2\right]\right] \\
&= p(\mathbf{x}) - \frac{1}{2}p(\mathbf{x})(1-p(\mathbf{x}))B_i + p(\mathbf{x})(1-p(\mathbf{x}))o\|\mathbf{x}_i-\mathbf{x}\|^2,
\end{aligned}$$

where $B_i = [\mathbf{x}_i-\mathbf{x}]^T\mathcal{H}(\mathbf{x})[\mathbf{x}_i-\mathbf{x}]$. Now,

$$\begin{aligned}
b^{''}(\mathbf{X}\beta_i) &= \frac{\exp(\mathbf{X}\beta_i)}{[1+\exp(\mathbf{X}\beta_i)]^2} \\
&= \frac{\exp(\mathbf{X}\beta_i)}{[1+\exp(\mathbf{X}\beta_i)]}\frac{1}{[1+\exp(\mathbf{x}\beta_i)]} \\
&= b^{'}(\mathbf{X}\beta_i)[1-b^{'}(\mathbf{X}\beta_i)] \\
&= p(\mathbf{x})[1-p(\mathbf{x})] + o\|\mathbf{x}_i-\mathbf{x}\|^2.
\end{aligned}$$

Since $E[\mathbf{y} - b^{'}(\mathbf{X}\beta)|\mathbf{X}_1,...,\mathbf{X}_n] = p(\mathbf{x})$,

$$E\left[\mathbf{y} - b^{'}(\mathbf{X}\beta)|\mathbf{X}_1,...,\mathbf{X}_n\right] \approx \begin{pmatrix} \frac{1}{2}p(\mathbf{x})(1-p(\mathbf{x}))B_1 \\ \vdots \\ \frac{1}{2}p(\mathbf{x})(1-p(\mathbf{x}))B_n \end{pmatrix}$$

Let $B(\mathbf{x}) = [B_1(\mathbf{x}),\ldots,B_n(\mathbf{x})]^T$ be the $n\times 1$ vector. Therefore,

$$E[y - b^{'}(\mathbf{X}\beta)|\mathbf{X}_1,...,\mathbf{X}_n] \approx \frac{1}{2}p(\mathbf{x})(1-p(\mathbf{x}))B(\mathbf{x}).$$

Then from equation (7.0.1),

$$E\left[\hat{\beta}(\mathbf{x})-\beta(\mathbf{x})|\mathbf{X}_1,...,\mathbf{X}_n\right]=\frac{1}{2}\mathbf{e}_1^T\left[\mathbf{X}_\mathbf{x}^T\text{diag}(b''(\mathbf{X}_\mathbf{x}\hat{\beta}))W_\mathbf{x}\mathbf{X}_\mathbf{x}\right]^{-1}\mathbf{X}_\mathbf{x}^TW_\mathbf{x}p(\mathbf{x})(1-p(\mathbf{x}))B(\mathbf{x}).$$

To compute the leading bias term for $\hat{\beta}$, let's derive $n^{-1}\mathbf{X}_\mathbf{x}^T\text{diag}(b''(\mathbf{X}_\mathbf{x}\hat{\beta}))W_\mathbf{x}\mathbf{X}_\mathbf{x}$ and $n^{-1}\mathbf{X}_\mathbf{x}^TW_\mathbf{x}p(\mathbf{x})(1-p(\mathbf{x}))B(\mathbf{x})$.

$$n^{-1}\mathbf{X}_\mathbf{x}^T\text{diag}\{b''(\mathbf{X}_\mathbf{x}\hat{\beta})\}W_\mathbf{x}\mathbf{X}_\mathbf{x}=\begin{pmatrix}1&\dots&1\\X_{11}-x_{01}&\dots&X_{1n}-x_{01}\\X_{21}-x_{02}&\dots&X_{2n}-x_{02}\end{pmatrix}\begin{pmatrix}p(1-p)&\cdots&0\\\vdots&\ddots&0\\0&\cdots&p(1-p)\end{pmatrix}$$
$$\begin{pmatrix}K_h(\mathbf{X}_1-\mathbf{x})&\cdots&0\\\vdots&\ddots&0\\0&\cdots&K_h(\mathbf{X}_n-\mathbf{x})\end{pmatrix}\begin{pmatrix}1&X_{11}-x_{01}&X_{21}-x_{02}\\\vdots&\vdots&\vdots\\1&X_{1n}-x_{01}&X_{2n}-x_{02}\end{pmatrix}.$$

After applying theory of matrix multiplication, this is simplified to $p(\mathbf{x})(1-p(\mathbf{x}))\times$

$$\begin{pmatrix}n^{-1}\sum_{i=1}^n\kappa_i&n^{-1}\sum_{i=1}^n\kappa_i(x_{i1}-x_{01})&n^{-1}\sum_{i=1}^n\kappa_i(x_{i2}-x_{02})\\n^{-1}\sum_{i=1}^n\kappa_i(x_{i1}-x_{01})&n^{-1}\sum_{i=1}^n\kappa_i(x_{i1}-x_{01})^2&n^{-1}\sum_{i=1}^n\kappa_i(x_{i1}-x_{01})(x_{i2}-x_{02})\\n^{-1}\sum_{i=1}^n\kappa_i(x_{i2}-x_{02})&n^{-1}\sum_{i=1}^n\kappa_i(x_{i1}-x_{01})(x_{i2}-x_{02})&n^{-1}\sum_{i=1}^n\kappa_i(x_{i2}-x_{02})^2\end{pmatrix}$$

Using standard results from density estimation (See Ruppert & Wand, 1994),

$$\begin{aligned}n^{-1}\sum_{i=1}^n\kappa_i&=n^{-1}\sum_{i=1}^nK_h(\mathbf{X}_i-\mathbf{x})\\&=\tilde{f}(\mathbf{x});\text{ where}\end{aligned}$$

$$\tilde{f}(\mathbf{x})=(n_1\hat{f}(\mathbf{x})+n_2\hat{g}(\mathbf{x}))/n$$

$$\begin{aligned}n^{-1}\sum_{i=1}^n\kappa_i(\mathbf{X}_i-\mathbf{x})&=n^{-1}\sum_{i=1}^nK_h(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})\\&=h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x})+o(h^2);\text{ where}\end{aligned}$$

$\mathcal{D}_{\tilde{f}}(\mathbf{x})$ denote the 2×1 vector of first order partial derivatives of $\tilde{f}$ and

$$\mathcal{D}_{\tilde{f}}(\mathbf{x}) = (n_1\mathcal{D}_f(\mathbf{x}) + n_2\mathcal{D}_g(\mathbf{x}))/n.$$

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\kappa_i(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T &= n^{-1}\sum_{i=1}^{n}K_h(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T \\ &= h^2\mu_2(K)\tilde{f}(\mathbf{x}) + o(h^2). \end{aligned}$$

Following these approximations, we get

$$\begin{aligned} (n^{-1}\mathbf{X}_{\mathbf{x}}^T\mathrm{diag}\{b''(\mathbf{X}_{\mathbf{x}}\hat{\beta})\}W_{\mathbf{x}}\mathbf{X}_{\mathbf{x}})^{-1} =& [p(\mathbf{x})(1-p(\mathbf{x}))]^{-1} \\ &\times \begin{pmatrix} \tilde{f}(\mathbf{x}) & h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) \\ h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) \\ h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) \end{pmatrix}^{-1}. \end{aligned}$$

Denote the above matrix M. Then we get,

$$(n^{-1}\mathbf{X}_{\mathbf{x}}^T\mathrm{diag}\{b''(\mathbf{X}_{\mathbf{x}}\hat{\beta})\}W_{\mathbf{x}}\mathbf{X}_{\mathbf{x}})^{-1} = [p(\mathbf{x})(1-p(\mathbf{x}))]^{-1} \times M^{-1},$$

where

$$M = \begin{pmatrix} \tilde{f}(\mathbf{x}) & h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) \\ h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) \\ h^2\mu_2(K)\mathcal{D}_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) \end{pmatrix}.$$

We use the theory of partitioned matrix to derive the inverse of M. Suppose

$$M = \begin{pmatrix} A_{2\times 2} & B_{2\times 1} \\ C_{1\times 2} & D_{1\times 1} \end{pmatrix}$$

where

$$A = \begin{pmatrix} \tilde{f}(\mathbf{x}) & h^2\mu_2(K)D_{\tilde{f}}(\mathbf{x}) \\ h^2\mu_2(K)D_{\tilde{f}}(\mathbf{x}) & h^2\mu_2(K)\tilde{f}(\mathbf{x}) \end{pmatrix},$$

$$B = \begin{pmatrix} h^2\mu_2(K)D_{\tilde{f}}(\mathbf{x}) \\ h^2\mu_2(K)\tilde{f}(\mathbf{x}) \end{pmatrix},$$

$$C = [h^2\mu_2(K)D_{\tilde{f}}(\mathbf{x}), h^2\mu_2(K)\tilde{f}(\mathbf{x})]^T$$

and

$$D = h^2\mu_2(K)\tilde{f}(\mathbf{x}).$$

Then

$$M^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1}.$$

According to the theory of partition matrices we get,

$$M^{-1} = \begin{pmatrix} A^{-1} + FE^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{pmatrix}$$

where $E = D - CA^{-1}B$ and $F = A^{-1}B$. Here we show that

$$A^{-1} = \begin{pmatrix} h^2\mu_2(K)\tilde{f}(\mathbf{x}) & -h^2\mu_2(K)D_{\tilde{f}}(\mathbf{x}) \\ -h^2\mu_2(K)D_{\tilde{f}}(\mathbf{x}) & \tilde{f}(\mathbf{x}) \end{pmatrix} \left[h^2\mu_2(K)\tilde{f}(\mathbf{x})^2 - h^4\mu_2(K)^2D^2_{\tilde{f}}(\mathbf{x})\right]^{-1}$$

$$= \begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} \\ -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & (h^2\mu_2(K)\tilde{f}(\mathbf{x}))^{-1} \end{pmatrix}$$

$$F = A^{-1}B = [0 \quad (1 - h^2\mu_2(K)^2D^2_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2})]_{1\times 2}$$

$$CA^{-1}B = h^2\mu_2(K)\tilde{f}(\mathbf{x})$$

and

$$E = D - CA^{-1}B = 0.$$

Therefore,

$$M^{-1} = \begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & 0 \\ -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & (h^2\mu_2(K)\tilde{f}(\mathbf{x}))^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{7.0.3}$$

From equation (7.0.1),

$$\begin{aligned} n^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} E[\mathbf{y} - b^{'}(\mathbf{X}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n] =& (n^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} \frac{1}{2} p(\mathbf{x})(1-p(\mathbf{x}))B(\mathbf{x}) \\ =& (n^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} \frac{1}{2} p(\mathbf{x})(1-p(\mathbf{x}))B(\mathbf{x}). \end{aligned}$$

$$\begin{aligned} n^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} E[y - b^{'}(\mathbf{X}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n] = \frac{1}{2} p(\mathbf{x})(1-p(\mathbf{x}))n^{-1} & \begin{pmatrix} 1 & \dots & 1 \\ (x_{11}-x_{01}) & \dots & (x_{1n}-x_{01}) \\ (x_{21}-x_{02}) & \dots & (x_{2n}-x_{02}) \end{pmatrix} \\ & \begin{pmatrix} \kappa_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \kappa_n \end{pmatrix} \begin{pmatrix} (\mathbf{X}_1-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{X}_1-\mathbf{x}) \\ \vdots \\ (\mathbf{X}_n-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{X}_n-\mathbf{x}) \end{pmatrix}. \end{aligned}$$

That is, $n^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} E[\mathbf{y} - b^{'}(\mathbf{X}_{\mathbf{x}}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n]$ is equal to $\frac{1}{2}p(\mathbf{x})(1-p(\mathbf{x}))\times$

$$\begin{pmatrix} n^{-1}\sum_{i=1}^n K_h(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{X}_i-\mathbf{x}) \\ n^{-1}\sum_{i=1}^n \{K_h(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\}(X_{1i}-x_{01}) \\ n^{-1}\sum_{i=1}^n \{K_h(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\}(X_{2i}-x_{02}) \end{pmatrix}.$$

Using Ruppert & Wand (1994), we get

$$n^{-1}\sum_{i=1}^n \{K_h(\mathbf{X}_i-\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T\mathcal{H}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\}(X_{1i}-x_{01}) \approx o(h^2).$$

Therefore, $n^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} E[\mathbf{y} - b'(\mathbf{X}_{\mathbf{x}}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n]$ is equal to $\frac{1}{2}p(\mathbf{x})(1-p(\mathbf{x}))\times$

$$\begin{pmatrix} n^{-1}\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T \mathcal{H}(\mathbf{x})(\mathbf{X}_i - \mathbf{x}) \\ o(h^2) \\ o(h^2) \end{pmatrix}.$$

By substituting the above results in equation (7.0.1),

$$\begin{aligned} E[\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x})] =& \frac{1}{2}\mathbf{e}_1^T [\mathbf{X}_{\mathbf{x}}^T \text{diag}(b''(\mathbf{X}_{\mathbf{x}}\hat{\beta}))W_{\mathbf{x}}\mathbf{X}_{\mathbf{x}}]^{-1}\mathbf{X}_{\mathbf{x}}^T W_{\mathbf{x}} p(\mathbf{x})(1-p(\mathbf{x}))B(\mathbf{x}) \\ =& \frac{1}{2}e_1^T [p(1-p)]^{-1} M^{-1}[p(1-p)] \begin{pmatrix} n^{-1}\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T \mathcal{H}(\mathbf{x})(\mathbf{X}_i - \mathbf{x}) \\ o(h^2) \\ o(h^2) \end{pmatrix} \\ =& \frac{1}{2}\tilde{f}(\mathbf{x})^{-1} n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^T \mathcal{H}(\mathbf{x})(\mathbf{X}_i - \mathbf{x}) \\ \approx& \frac{1}{2}\tilde{f}(\mathbf{x})^{-1} \int\int K_h(\mathbf{y} - \mathbf{x})(\mathbf{y} - \mathbf{x})^T \mathcal{H}(\mathbf{x})(\mathbf{y} - \mathbf{x})\tilde{f}(\mathbf{y})d\mathbf{y}. \end{aligned}$$

By substituting $\mathbf{z} = \frac{\mathbf{y}-\mathbf{x}}{h}$, we get

$$\begin{aligned} E[\hat{\beta}(\mathbf{x}) - \beta(\mathbf{x})] =& \frac{1}{2}\tilde{f}(\mathbf{x})^{-1} \int\int \frac{1}{h^2} K(\mathbf{z})(\mathbf{zh})^T \mathcal{H}(\mathbf{x})(\mathbf{zh})\tilde{f}(\mathbf{x} + h\mathbf{z})h^2 d\mathbf{z} \\ =& \frac{1}{2}\tilde{f}(\mathbf{x})^{-1}\tilde{f}(\mathbf{x}) \int\int h^T \mathcal{H}(\mathbf{x}) h \{K(\mathbf{z})\mathbf{z}\mathbf{z}^T\} d\mathbf{z} \\ =& \frac{1}{2} tr[h^T \mathcal{H}(\mathbf{x}) h \int\int K(\mathbf{z})\mathbf{z}\mathbf{z}^T \mathbf{dz}] \\ =& \frac{1}{2}h^2 \mu_2(K) tr\mathcal{H}(\mathbf{x}) \\ =& \frac{1}{2}h^2 \mu_2(K) \left[\frac{\partial^2}{\partial x_1^2}\beta(\mathbf{x}) + \frac{\partial^2}{\partial x_2^2}\beta(\mathbf{x})\right]. \end{aligned}$$

Recall that the local linear log relative risk estimator, $\breve{\rho}_{LL} = \beta_0 - \text{Constant}$, where $\text{Constant} = \log(\frac{n_1}{n_2})$. Therefore,

$$E[\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})] = \frac{1}{2}h^2 \mu_2(K) \left[\frac{\partial^2}{\partial x_1^2}\rho(\mathbf{x}) + \frac{\partial^2}{\partial x_2^2}\rho(\mathbf{x})\right]$$

$$E[\hat{\rho}(\mathbf{x}) - \rho(\mathbf{x})] = \frac{1}{2}h^2\mu_2(K)\nabla^2\rho(\mathbf{x})$$

as required, where ∇^2 is the Laplacian operator.

Now we obtain the approximation to the variance term. Note that

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) = \mathbf{e}_1^T \Big[&\left(\mathbf{X}_\mathbf{x}^T \mathrm{diag}(b^{"}(\mathbf{X}_\mathbf{x}\beta))W_\mathbf{x}\mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^T W_\mathbf{x} \mathrm{Var}[\mathbf{y} - b^{'}(\mathbf{X}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n] \\ &W_\mathbf{x}\mathbf{X}_\mathbf{x} \left(\mathbf{X}_\mathbf{x}^T \mathrm{diag}(b^{"}(\mathbf{X}_\mathbf{x}\beta))W_\mathbf{x}\mathbf{X}_\mathbf{x}\right)^{-1}\Big] \mathbf{e}_1.\end{aligned}$$

We first derive $n^{-1}\mathbf{X}_\mathbf{x}^T W_\mathbf{x} \mathrm{Var}[\mathbf{y} - b^{'}(\mathbf{X}_\mathbf{x}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n] W_\mathbf{x}\mathbf{X}_\mathbf{x}$ (3×3 matrix).

Since $\mathrm{Var}\left[\mathbf{y} - b^{'}(\mathbf{X}_\mathbf{x}\beta)|\mathbf{X}_1, ..., \mathbf{X}_n\right] = \mathrm{diag}\{p(\mathbf{x}_1)(1-p(\mathbf{x}_1)), \ldots, p(\mathbf{x}_n)(1-p(\mathbf{x}_n))\}$ $(= V)$,

where $V(\mathbf{x}_i) = p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))$.

$$\begin{aligned}n^{-1}\mathbf{X}_\mathbf{x}^T W_\mathbf{x} V W_\mathbf{x}\mathbf{X}_\mathbf{x} = n^{-1} &\begin{pmatrix} 1 & \ldots & 1 \\ (X_{11} - x_{01}) & \ldots & (X_{1n} - x_{01}) \\ (X_{21} - x_{02}) & \ldots & (X_{2n} - x_{02}) \end{pmatrix} \\ &\begin{pmatrix} K_h(\mathbf{X}_1 - \mathbf{x}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_h(\mathbf{X}_n - \mathbf{x}) \end{pmatrix} \begin{pmatrix} v(x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v(x_n) \end{pmatrix} \\ &\begin{pmatrix} K_h(\mathbf{X}_1 - \mathbf{x}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_h(\mathbf{X}_n - \mathbf{x}) \end{pmatrix} \begin{pmatrix} 1 & (X_{11} - x_{01}) & (X_{21} - x_{02}) \\ \vdots & \vdots & \vdots \\ 1 & (X_{1n} - x_{01}) & (X_{2n} - x_{02}) \end{pmatrix}.\end{aligned}$$

After using matrix multiplication several times, $n^{-1}\mathbf{X}_\mathbf{x}^T W_\mathbf{x} V W_\mathbf{x}\mathbf{X}_\mathbf{x}$ reaches to the following single matrix.

$$\begin{pmatrix} \Sigma_{i=1}^n n^{-1}\kappa_i^2 v(x_i) & \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i1} - x_{01})v(x_i) & \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i2} - x_{02})v(x_i) \\ \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i1} - x_{01})v(x_i) & \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i1} - x_{01})^2 v(x_i) & \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i1} - x_{01})(x_{i2} - x_{02})v(x_i) \\ \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i2} - x_{02})v(x_i) & \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i1} - x_{01})(x_{i2} - x_{02})v(x_i) & \Sigma_{i=1}^n n^{-1}\kappa_i^2 (x_{i2} - x_{02})^2 v(x_i). \end{pmatrix}.$$

For the convenience, we label the components of this by c_{ij}; where i represents the row number while j, the column number. Then,

$$\begin{aligned}
c_{11} =& \sum_{i=1}^{n} n^{-1}\kappa_i^2 v(\mathbf{x}_i) \\
\approx& \int\int \frac{1}{h^4} K(\frac{\mathbf{y}-\mathbf{x}}{h})^2 v(\mathbf{y})\tilde{f}(\mathbf{y})d\mathbf{y} \\
=& \int\int \frac{1}{h^2} K(\mathbf{z})^2 \{v(\mathbf{x}) + (h\mathbf{z})^T \mathcal{D}_v(\mathbf{z}) + \ldots\}\{\tilde{f}(\mathbf{x}) + (h\mathbf{z})^T \mathcal{D}_{\tilde{f}}(\mathbf{x}) + \ldots\} d\mathbf{z} \\
=& h^{-2} v(\mathbf{x})\tilde{f}(\mathbf{x}) R(K) + o(h^2) \\
\approx& h^{-2} v(\mathbf{x})\tilde{f}(\mathbf{x}) R(K)
\end{aligned}$$

$$\begin{aligned}
c_{12} =& \sum_{i=1}^{n} n^{-1}\kappa_i^2 v(x_i)(x_{i1} - x_{01}) \\
\approx& \int\int \frac{1}{h^4} K(\frac{\mathbf{y}-\mathbf{x}}{h})^2 v(\mathbf{y})(y_1 - x_{01})\tilde{f}(\mathbf{y})d\mathbf{y} \\
=& \int\int \frac{1}{h^2} K(\mathbf{z})^2 \{v(\mathbf{x}) + (h\mathbf{z})^T \mathcal{D}_v(\mathbf{z}) + \ldots\}(hz_1)\{f(\mathbf{x}) + (h\mathbf{z})^T \mathcal{D}_{\tilde{f}}(\mathbf{x}) + \ldots\} d\mathbf{z} \\
=& \int\int K(\mathbf{z})^2 h^{-1} z_1^2 v(\mathbf{x}) f(\mathbf{x}) + o(h^2) \\
\approx& o(h^2)
\end{aligned}$$

Similarly we can show that c_{13}, c_{21} and c_{31} are same as c_{12}.

$$\begin{aligned}
c_{22} =& \sum_{i=1}^{n} n^{-1}\kappa_i^2 v(x_i)(x_{i1} - x_{01})^2 \\
\approx& \int\int \frac{1}{h^4} K(\frac{\mathbf{y}-\mathbf{x}}{h})^2 v(y)(y_1 - x_{01})^2\tilde{f}(\mathbf{y})d\mathbf{y} \\
=& \int\int \frac{1}{h^2} K(\mathbf{z})^2 \{v(\mathbf{x}) + (h\mathbf{z})^T \mathcal{D}_v(\mathbf{z}) + \ldots\}(hz_1)^2\{f(x) + (h\mathbf{z})^T \mathcal{D}_{\tilde{f}}(\mathbf{x}) + \ldots\} d\mathbf{z} \\
=& \int\int K(\mathbf{z})^2 z_1^2 v(\mathbf{x}) f(\mathbf{x}) + o(h^2) \\
\approx& o(h^2)
\end{aligned}$$

Similarly we can show that $c_{33} = c_{22}$.

$$\begin{aligned}
c_{23} =& \sum_{i=1}^{n} n^{-1}\kappa_i^2 v(x_i)(x_{i1} - x_{01})(x_{i2} - x_{02}) \\
\approx& \int\int \frac{1}{h^4} K(\frac{\mathbf{y}-\mathbf{x}}{h})^2 v(\mathbf{y})(y_1 - x_{01})(y_2 - x_{02})\tilde{f}(\mathbf{y})d\mathbf{y} \\
=& \int\int \frac{1}{h^2} K(\mathbf{z})^2 \{v(\mathbf{x}) + (h\mathbf{z})^T \mathcal{D}_v(\mathbf{z}) + \ldots\}(hz_1)^2 \{f(\mathbf{x}) + (\mathbf{hz})^T \mathcal{D}_{\tilde{f}}(\mathbf{x}) + \ldots\} d\mathbf{z} \\
=& \int\int K(\mathbf{z})^2 z_1 z_2 v(\mathbf{x}) f(\mathbf{x}) d\mathbf{z} + o(h^2) \\
\approx& o(h^2)
\end{aligned}$$

By symmetry, $c_{32} = c_{23}$. Therefore,

$$n^{-1}\mathbf{X}_\mathbf{x}^T W_\mathbf{x} V W_\mathbf{x} \mathbf{X}_\mathbf{x} = \begin{pmatrix} h^{-2} v(\mathbf{x})\tilde{f}(\mathbf{x}) R(K) & o(h^2) & o(h^2) \\ o(h^2) & o(h^2) & o(h^2) \\ o(h^2) & o(h^2) & o(h^2) \end{pmatrix}$$

By using these results we obtained the approximation to the variance.

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) =& n^{-1}\{p(\mathbf{x})(1-p(\mathbf{x}))\}^{-2}\mathbf{e}_1^T
\begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & 0 \\ -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & (h^2\mu_2(K)\tilde{f}(\mathbf{x}))^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}
\begin{pmatrix} h^{-2}v(\mathbf{x})\tilde{f}(\mathbf{x})R(K) & op(h^2) & op(h^2) \\ op(h^2) & o(h^2) & o(h^2) \\ op(h^2) & o(h^2) & o(h^2) \end{pmatrix} \\
& \begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & 0 \\ -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & (h^2\mu_2(K)\tilde{f}(\mathbf{x}))^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}\mathbf{e}_1 + \dots \\
=& n^{-1}\{p(\mathbf{x})(1-p(\mathbf{x}))\}^{-2}
\begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & 0 \end{pmatrix}
\begin{pmatrix} h^{-2}v(\mathbf{x})\tilde{f}(\mathbf{x})R(K) & op(h^2) & op(h^2) \\ op(h^2) & op(h^2) & op(h^2) \\ op(h^2) & op(h^2) & op(h^2) \end{pmatrix} \\
& \begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & 0 \\ -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & (h^2\mu_2(K)\tilde{f}(\mathbf{x}))^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}e_1 + \dots \\
=& n^{-1}\{p(\mathbf{x})(1-p(\mathbf{x}))\}^{-2}
\begin{pmatrix} h^{-2}v(\mathbf{x})R(K) & o(h^2) & o(h^2) \end{pmatrix}
\begin{pmatrix} \tilde{f}(\mathbf{x})^{-1} & -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & 0 \\ -D_{\tilde{f}}(\mathbf{x})\tilde{f}(\mathbf{x})^{-2} & (h^2\mu_2(K)\tilde{f}(\mathbf{x}))^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}e_1 + \dots \\
=& n^{-1}\{p(\mathbf{x})(1-p(\mathbf{x}))\}^{-2}h^{-2}v(\mathbf{x})R(K)\tilde{f}(\mathbf{x})^{-1} + \dots \\
=& n^{-1}[p(\mathbf{x})(1-p(\mathbf{x}))]^{-2}h^{-2}[p(\mathbf{x})(1-p(\mathbf{x}))]R(K)\frac{n}{n_1 f(\mathbf{x}) + n_2 g(\mathbf{x})} + \dots \\
=& h^{-2}R(K)[p(\mathbf{x})(1-p(\mathbf{x}))]^{-1}\frac{1}{n_1 f(\mathbf{x}) + n_2 g(\mathbf{x})} + \dots \\
=& h^{-2}R(K)\frac{1}{n_1 f(\mathbf{x}) + n_2 g(\mathbf{x})}\frac{[n_1 f(\mathbf{x}) + n_2 g(\mathbf{x})]^2}{n_1 f(\mathbf{x})\ n_2 g(\mathbf{x})} + \dots \\
=& h^{-2}R(K)\left[\frac{n_1 f(\mathbf{x}) + n_2 g(\mathbf{x})}{n_1 f(\mathbf{x})\ n_2 g(\mathbf{x})}\right] + \dots \\
=& h^{-2}R(K)\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right] + \dots .
\end{aligned}
$$

Since $\breve{\rho}_{LL} = \hat{\beta}_0 - \text{Constant}$,

$$\mathrm{Var}(\breve{\rho}_{LL}) = h^{-2}R(K)\left[\frac{1}{n_1 f(\mathbf{x})} + \frac{1}{n_2 g(\mathbf{x})}\right] + o(n_1^{-1}h^{-2} + n_2^{-1}h^{-2}).$$

□

C.1: Proof of Theorem (5.5.2)

Proof. As

$$\hat{\rho}(\mathbf{z};t) = \log\left[\frac{\hat{f}(\mathbf{z};t)|T|}{\hat{g}(\mathbf{z})}\right] \tag{7.0.4}$$

Let us define the relative error term, $\epsilon_f(\mathbf{z};t)$ of $\hat{f}(\mathbf{z};t)$

$$\epsilon_f(\mathbf{z};t) = \frac{\hat{f}(\mathbf{z};t) - f(\mathbf{z};t)}{f(\mathbf{z};t)}.$$

Therefore,

$$\hat{f}(\mathbf{z};t) = f(\mathbf{z};t)\{1+\epsilon_f(\mathbf{z};t)\}.$$

Similarly,

$$\hat{g}(\mathbf{z}) = g(\mathbf{z})\{1+\epsilon_g(\mathbf{z})\}.$$

Therefore the equation (7.0.4) becomes,

$$\begin{aligned}
\hat{\rho}(\mathbf{z};t) &= \log\left[\frac{f(\mathbf{z};t)\{1+\epsilon_f(\mathbf{z};t)\}|T|}{g(\mathbf{z})\{1+\epsilon_g(\mathbf{z})\}}\right] \\
&= \rho(\mathbf{z}) + \log\{1+\epsilon_f(\mathbf{z};t)\} - \log\{1+\epsilon_g(\mathbf{z})\} \\
&= \rho(\mathbf{z};t) + \epsilon_f(\mathbf{z};t) - \epsilon_g(\mathbf{z}) + o(\epsilon_f^2+\epsilon_g^2).
\end{aligned}$$

Assuming that the error terms are small, the expectation of $\hat{\rho}(\mathbf{z};t)$ gives

$$E[\hat{\rho}(\mathbf{z};t)] \approx \rho(\mathbf{z};t) + E[\epsilon_f(\mathbf{z};t)] - E[\epsilon_g(\mathbf{z})] \tag{7.0.5}$$

where

$$\begin{aligned}
E[\epsilon_f(\mathbf{z};t)] &= \frac{E[\hat{f}(\mathbf{z};t) - f(\mathbf{z};t)]}{f(\mathbf{z};t)} \\
&= \frac{\text{Bias}[\hat{f}(\mathbf{z};t)]}{f(\mathbf{z};t)} \\
&= \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}f(\mathbf{z};t) + \lambda^2\mu_2(L)\nabla^2_t f(\mathbf{z};t)}{f(\mathbf{z};t)}\right] + o(h^2+\lambda^2).
\end{aligned}$$

Similarly,

$$E[\epsilon_g(\mathbf{z})] = \left[\frac{1}{2}h^2\mu_2(K)\frac{\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})}\right] + o(h^2).$$

Therefore from equation (7.0.5),

$$\begin{aligned} E[\hat{\rho}(\mathbf{z};t)] &= \rho(\mathbf{z};t) + \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}f(\mathbf{z};t) + \lambda^2\mu_2(L)\nabla^2_t(\mathbf{z};t)}{f(\mathbf{z};t)} - \frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})}\right] \\ &\quad +o(h^2+\lambda^2). \end{aligned}$$

The variance term of $\hat{\rho}$ is

$$\begin{aligned} \mathrm{Var}[\hat{\rho}(\mathbf{z};t)] &= \mathrm{Var}(\epsilon_f) + \mathrm{Var}(\epsilon_g) + \mathrm{Cov}(\epsilon_f, \epsilon_g) \\ &= \frac{\mathrm{Var}\{\hat{f}(\mathbf{z};t)\}}{f(\mathbf{z};t)^2} + \frac{\mathrm{Var}\{\hat{g}(\mathbf{z})\}}{g(\mathbf{z})^2}(\text{Error terms are independent.}) \\ &= \frac{R(K)R(L)}{n_1h^2\lambda f(\mathbf{z};t)} + \frac{R(K)}{n_2h^2g(\mathbf{z})} + o(n_1^{-1}h^{-2}\lambda^{-1} + n_2^{-1}h^{-2}). \end{aligned}$$

In conclusion,

$$\begin{aligned} \mathrm{Bias}[\hat{\rho}(\mathbf{z};t)] &= \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}f(\mathbf{z};t) + \lambda^2\mu_2(L)\nabla^2_t f(\mathbf{z};t)}{f(\mathbf{z};t)} - \frac{h^2\mu_2(K)\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})}\right] \\ &\quad +o(h^2+\lambda^2) \end{aligned}$$

and

$$\mathrm{Var}[\hat{\rho}(\mathbf{z};t)] = \frac{R(K)R(L)}{n_1h^2\lambda}\{f(\mathbf{z};t)\}^{-1} + \frac{R(K)}{n_2h^2}\{g(\mathbf{z})\}^{-1} + o(n_1^{-1}h^{-2}\lambda^{-1} + n_2^{-1}h^{-2}).$$

□

C.2: Proof of Theorem (5.9.1)

Proof. By using the second order 'delta method', we get

$$\begin{aligned} E\left[\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)\right] &= E\left[\frac{\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)}{\hat{f}(\mathbf{z};t)}\right] \\ &\approx \frac{\frac{\partial}{\partial t}f(\mathbf{z};t)}{f(\mathbf{z};t)} - \frac{\mathrm{Cov}[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t),\hat{f}(\mathbf{z};t)]}{[f(\mathbf{z},t)]^2} + \frac{\mathrm{Var}[\hat{f}(\mathbf{z},t)]\frac{\partial}{\partial t}f(\mathbf{z};t)}{[f(\mathbf{z},t)]^3}. \end{aligned} \tag{7.0.6}$$

Recall from equation (5.5.2),

$$\mathrm{Var}[\hat{f}(\mathbf{z},t)] = \frac{1}{n_1h^2\lambda}R(K)R(L)f(\mathbf{z};t) + o\left(\frac{1}{n_1h^2\lambda}\right).$$

To derive the asymptotic bias, we first need to derive $\mathrm{Cov}\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t),\hat{f}(\mathbf{z};t)\right]$ term.

$$\begin{aligned} \mathrm{Cov}\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t),\hat{f}(z;t)\right] &= E\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t).\hat{f}(\mathbf{z};t)\right] - E\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)\right]E\left[\hat{f}(\mathbf{z};t)\right] \\ &= \frac{1}{2}E\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)^2\right] - \frac{1}{2}\frac{\partial}{\partial t}(E[\hat{f}(\mathbf{z};t)])^2 \\ &= \frac{1}{2}\frac{\partial}{\partial t}\left[\mathrm{Var}(\hat{f}(\mathbf{z};t))\right] \\ &= \frac{1}{2}\frac{\partial}{\partial t}\left[\frac{1}{n_1h^2\lambda}R(K)R(L)f(\mathbf{z};t)\right] \\ &= \frac{R(K)R(L)}{2n_1h^2\lambda}\ \frac{\partial}{\partial t}f(\mathbf{z};t) + o\left(\frac{1}{n_1h^2\lambda}\right). \end{aligned}$$

By substituting equations (7.0.7) and (5.5.2) in (7.0.6), we get

$$E\left[\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)\right] = \frac{\frac{\partial}{\partial t}f(\mathbf{z};t)}{f(\mathbf{z};t)} + \frac{R(K)R(L)\frac{\partial}{\partial t}f(\mathbf{z};t)}{2n_1h^2\lambda\ [f(\mathbf{z},t)]^2} + o(n_1^{-1}h^{-2}\lambda^{-1}).$$

Let us find the variance of $\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)$, which is equal to $\mathrm{Var}\left[\frac{\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)}{\hat{f}(\mathbf{z};t)}\right]$. Here, the derivation of the variance of a ratio $\frac{X_1}{X_2}$ is somewhat difficult. So we use first order

'delta method' to derive the formula for the variance of a ratio $\frac{X_1}{X_2}$ which is given as follows.

$$\text{Var}\left[\frac{X_1}{X_2}\right] = \left(\frac{\mu_1}{\mu_2}\right)^2 \times \left[\frac{\text{Var}(X_1)}{\mu_1^2} - \frac{2\text{Cov}(X_1, X_2)}{\mu_1\mu_2} + \frac{\text{Var}(X_2)}{\mu_2^2}\right]$$

where $\mu_1 = E[X_1]$ and $\mu_2 = E[X_2]$. So in particular,

$$\begin{aligned}\text{Var}\left[\frac{\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)}{\hat{f}(\mathbf{z};t)}\right] \approx & \left[\frac{\frac{\partial}{\partial t}f(\mathbf{z};t)}{f(\mathbf{z};t)}\right]^2 \times \\ & \left[\frac{\text{Var}[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)]}{(\frac{\partial}{\partial t}f(\mathbf{z};t))^2} - \frac{2\text{Cov}[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t), \hat{f}(\mathbf{z};t)]}{(\frac{\partial}{\partial t}f(\mathbf{z};t))f(\mathbf{z};t)} + \frac{\text{Var}[\hat{f}(\mathbf{z};t)]}{(f(\mathbf{z};t))^2}\right].\end{aligned} \tag{7.0.7}$$

In order to derive this asymptotic variance formulae, we need first to obtain the approximation for $\text{Var}[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)]$.

$$\begin{aligned}\text{Var}\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)\right] &= \text{Var}\left[\frac{\partial}{\partial t}\sum_i \frac{1}{n_1}K(\mathbf{z}-\mathbf{x}_i;t-t_i)\right] \\ &= \frac{1}{n_1}\text{Var}\left[\frac{\partial}{\partial t}\sum_i K(\mathbf{z}-\mathbf{x}_i;t-t_i)\right] \\ &= \frac{1}{n_1}\int\int\left(\frac{\partial}{\partial t}K_h(\mathbf{z}-\mathbf{y})L_\lambda(t-s)\right)^2 f(\mathbf{y})f(s)d\mathbf{y}ds + o(n_1^{-1}) \\ &= \frac{1}{n_1 h^4\lambda^4}\int\int K\left(\frac{\mathbf{z}-\mathbf{y}}{h}\right)^2 L'\left(\frac{t-s}{\lambda}\right)^2 f(\mathbf{y})f(s)d\mathbf{y}ds.\end{aligned}$$

By substituting $\mathbf{u} = (\mathbf{z}-\mathbf{y})/h$, $v = (t-s)/\lambda$ and applying first order Taylor expansion, we get

$$\begin{aligned}\text{Var}\left[\frac{\partial}{\partial t}\hat{f}(\mathbf{z};t)\right] &= \frac{1}{n_1 h^2\lambda^3}\int\int K(\mathbf{u})^2(L'(v))^2 f(\mathbf{z},t)d\mathbf{u}dv + o(n_1^{-1}h^{-2}\lambda^{-3}). \\ &= \frac{1}{n_1 h^2\lambda^3}R(K)R(L')f(\mathbf{z};t) + o(n_1^{-1}h^{-2}\lambda^{-3}).\end{aligned}$$

Substituting equations (7.0.7), (5.5.2) and (7.0.8) in (7.0.8),

$$\begin{aligned}
\text{Var}\left[\frac{\partial}{\partial t}\hat{\rho}(\mathbf{z};t)\right] &\approx \left(\frac{\frac{\partial}{\partial t}f(\mathbf{z};t)}{f(\mathbf{z};t)}\right)^2 \times \left[\frac{1}{(\frac{\partial}{\partial t}f(\mathbf{z};t))^2}\frac{1}{n_1h^2\lambda^3}R(K)R(L^{'})f(\mathbf{z};t)\right. \\
&\left. -\frac{1}{n_1h^2\lambda}R(K)R(L)\{f(\mathbf{z};t)\}^{-1} + \frac{1}{n_1h^2\lambda}\ R(K)R(L)f(\mathbf{z};t)\right] \\
&= \frac{R(K)R(L^{'})}{n_1h^2\lambda^3 f(\mathbf{z};t)} - \frac{R(K)R(L)}{n_1h^2\lambda}\frac{[\frac{\partial}{\partial t}f(\mathbf{z};t)]^2}{[f(\mathbf{z};t)]^3} + \frac{R(K)R(L)[\frac{\partial}{\partial t}f(\mathbf{z};t)]^2}{n_1h^2\lambda f(\mathbf{z};t)} \\
&+o\left(\frac{1}{n_1h^2\lambda} + \frac{1}{n_1h^2\lambda^3}\right) \\
&= \frac{R(K)R(L^{'})}{n_1h^2\lambda^3 f(\mathbf{z};t)} + o\left(\frac{1}{n_1h^2\lambda} + \frac{1}{n_1h^2\lambda^3}\right).
\end{aligned}$$

□

D.1: Proof of Lemma 6.4.1

Proof. From equation (6.2.3),

$$\begin{aligned}
E[\hat{f}(\mathbf{x}|\mathbf{z})] &= E\left[\frac{\hat{f}(\mathbf{x},\mathbf{z})}{\hat{f}(\mathbf{z})}\right] \\
&= \frac{f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})} + E\left[\frac{\hat{f}(\mathbf{x},\mathbf{z})}{\hat{f}(\mathbf{z})} - \frac{f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})}\right] \\
&= f(\mathbf{x}|\mathbf{z}) + E\left[\frac{\hat{f}(\mathbf{x},\mathbf{z})}{\hat{f}(\mathbf{z})} - \frac{f(\mathbf{x},\mathbf{z})}{\hat{f}(\mathbf{z})} + \frac{f(\mathbf{x},\mathbf{z})}{\hat{f}(\mathbf{z})} - \frac{f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})}\right] \\
&= f(\mathbf{x}|\mathbf{z}) + E\left[\frac{\text{Bias}(\hat{f}(\mathbf{x},\mathbf{z}))}{\hat{f}(\mathbf{z})} + f(\mathbf{x},\mathbf{z})\text{Bias}\left(\frac{1}{\hat{f}(\mathbf{z})}\right)\right] \\
&= f(\mathbf{x}|\mathbf{z}) + \frac{\text{Bias}(\hat{f}(\mathbf{x},\mathbf{z}))}{f(\mathbf{z})}(1+o(1)) + f(\mathbf{x},\mathbf{z})\text{Bias}\left(\frac{1}{\hat{f}(\mathbf{z})}\right) \\
&= f(\mathbf{x}|\mathbf{z}) + \frac{\text{Bias}(\hat{f}(\mathbf{x},\mathbf{z}))}{f(\mathbf{z})} - \frac{f(\mathbf{x},\mathbf{z})\text{Bias}(\hat{f}(\mathbf{z}))}{f(\mathbf{z})^2}.
\end{aligned}$$

By substituting the results of 2.4.4 and 6.4.1, we get

$$\begin{aligned}\text{Bias}[\hat{f}(\mathbf{x}|\mathbf{z})] &= \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})} + \frac{\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})} - \frac{\lambda^2\mu_2(L)f(\mathbf{x},\mathbf{z})\nabla^2_{\mathbf{z}}f(\mathbf{z})}{f(\mathbf{z})^2}\right] \\ &\quad + o(h^2+\lambda^2)\end{aligned}$$

Let us obtain the formulae for $\text{Var}[\hat{f}(\mathbf{x}|\mathbf{z})]$.

$$\begin{aligned}\text{Var}[\hat{f}(\mathbf{x}|\mathbf{z})] &= E[\hat{f}(\mathbf{x}|\mathbf{z})^2] - \{E[\hat{f}(\mathbf{x}|\mathbf{z})]\}^2 \\ &\approx \frac{1}{n_1 h^4 \lambda^{2p}} \int\int \frac{K(\mathbf{u})^2 L(\mathbf{v})^2 f(\mathbf{x}-\mathbf{u}h, \mathbf{z}-\mathbf{v}\lambda)}{f(\mathbf{z})^2} h^2\lambda^p \, d\mathbf{u}\, d\mathbf{v} \\ &\quad \text{(by applying second order Taylor expansion to } f(\mathbf{x}-h\mathbf{u}, \mathbf{z}-l\mathbf{v}) \\ &\quad \text{and simplifying)} \\ &= \frac{R(K)R(L)}{n_1 h^2\lambda^p}\frac{f(\mathbf{x}|\mathbf{z})}{f(\mathbf{z})} + o\left(\frac{1}{n_1 h^2\lambda^p}\right).\end{aligned}$$

□

D.2: Proof of Theorem 6.4.2

Proof.

$$\hat{\rho}(\mathbf{x}|\mathbf{z}) = \log\left[\frac{\hat{f}(\mathbf{x}|\mathbf{z})}{\hat{g}(\mathbf{x}|\mathbf{z})}\right].$$

We define the relative error term of $\hat{f}(\mathbf{x}|\mathbf{z})$ be ε_f and similarly for ε_g as follows.

$$\varepsilon_f = \frac{\hat{f}(\mathbf{x}|\mathbf{z}) - f(\mathbf{x}|\mathbf{z})}{f(\mathbf{x}|\mathbf{z})}$$

and

$$\varepsilon_g = \frac{\hat{g}(\mathbf{x}|\mathbf{z}) - g(\mathbf{x}|\mathbf{z})}{g(\mathbf{x}|\mathbf{z})}.$$

$$\therefore\ \hat{\rho}(\mathbf{x}|\mathbf{z}) = \rho(\mathbf{x}|\mathbf{z}) + \log\left(\frac{1+\varepsilon_f}{1+\varepsilon_g}\right)$$

(By applying first order Taylor's expansion).

By assuming the error terms are small,

$$\begin{aligned} E[\hat{\rho}(\mathbf{x}|\mathbf{z})] &\approx \rho(\mathbf{x}|\mathbf{z}) + E[\varepsilon_f] - E[\varepsilon_g] \\ &\approx \rho(\mathbf{x}|\mathbf{z}) + \frac{\text{Bias}[\hat{f}(\mathbf{x}|\mathbf{z})]}{f(\mathbf{x}|\mathbf{z})} - \frac{\text{Bias}[\hat{g}(\mathbf{x}|\mathbf{z})]}{g(\mathbf{x}|\mathbf{z})}. \end{aligned}$$

That is,

$$\text{Bias}[\hat{\rho}(\mathbf{x}|\mathbf{z})] \approx \frac{\text{Bias}[\hat{f}(\mathbf{x}|\mathbf{z})]}{f(\mathbf{x}|\mathbf{z})} - \frac{\text{Bias}[\hat{g}(\mathbf{x}|\mathbf{z})]}{g(\mathbf{x}|\mathbf{z})}.$$

By substituting the equation (6.4.2) and using the formulae for the bias of $\hat{g}$ and after some simplifications,

$$\begin{aligned} \text{Bias}[\hat{\rho}(\mathbf{x}|\mathbf{z})] &= \frac{1}{2f(\mathbf{x}|\mathbf{z})}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})} + \frac{\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{z})} - \frac{\lambda^2\mu_2(L)f(\mathbf{x},\mathbf{z})\nabla^2_{\mathbf{z}}f(\mathbf{z})}{f(\mathbf{z})^2}\right] \\ &\quad -\frac{1}{2g(\mathbf{x}|\mathbf{z})}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{z})} + \frac{\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{z})} - \frac{\lambda^2\mu_2(L)g(\mathbf{x},\mathbf{z})\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})^2}\right] \\ &\quad +o(h^2+\lambda^2) \\ &= \frac{1}{2}\left[\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{x},\mathbf{z})} + \frac{\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{x},\mathbf{z})} - \frac{\lambda^2\mu_2(K)\nabla^2_{\mathbf{z}}f(\mathbf{z})}{f(\mathbf{z})}\right. \\ &\quad \left.-\frac{h^2\mu_2(K)\nabla^2_{\mathbf{x}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{x},\mathbf{z})} - \frac{\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{x},\mathbf{z})} + \frac{\lambda^2\mu_2(L)\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})}\right] \\ &\quad +o(h^2+\lambda^2) \\ &= \frac{1}{2}h^2\mu_2(K)\left[\frac{\nabla^2_{\mathbf{x}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{x},\mathbf{z})} - \frac{\nabla^2_{\mathbf{x}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{x},\mathbf{z})}\right] + \frac{1}{2}\lambda^2\mu_2(L)\left[\frac{\nabla^2_{\mathbf{z}}f(\mathbf{x},\mathbf{z})}{f(\mathbf{x},\mathbf{z})} - \frac{\nabla^2_{\mathbf{z}}g(\mathbf{x},\mathbf{z})}{g(\mathbf{x},\mathbf{z})}\right] \\ &\quad -\frac{1}{2}\lambda^2\mu_2(L)\left[\frac{\nabla^2_{\mathbf{z}}f(\mathbf{z})}{f(\mathbf{z})} - \frac{\nabla^2_{\mathbf{z}}g(\mathbf{z})}{g(\mathbf{z})}\right] \\ &\quad +o(h^2+\lambda^2). \end{aligned}$$

By assuming the error terms are independent,

$$\begin{aligned}
\mathrm{Var}[\hat{\rho}(\mathbf{x}|\mathbf{z})] &\approx \mathrm{Var}(\varepsilon_f) + \mathrm{Var}(\varepsilon_g) \\
&\approx \frac{\mathrm{Var}[\hat{f}(\mathbf{x}|\mathbf{z})]}{f(\mathbf{x}|\mathbf{z})^2} + \frac{\mathrm{Var}[\hat{g}(\mathbf{x}|\mathbf{z})]}{g(\mathbf{x}|\mathbf{z})^2} \\
&= R(K)R(L)\left[\frac{1}{n_1 h^2 f(\mathbf{z}) f(\mathbf{x}|\mathbf{z})} + \frac{1}{n_2 \lambda^p g(\mathbf{z}) g(\mathbf{x}|\mathbf{z})}\right] + o\left(\frac{1}{n_1 h^2} + \frac{1}{n_2 \lambda^p}\right) \\
&= R(K)R(L)\left[\frac{1}{n_1 h^2 f(\mathbf{x}, \mathbf{z})} + \frac{1}{n_2 \lambda^p g(\mathbf{x}, \mathbf{z})}\right] + o\left(\frac{1}{n_1 h^2} + \frac{1}{n_2 \lambda^p}\right)
\end{aligned}$$

□

Bibliography

Abellan, J.J., Richardson, S. and Best, N. (2008). Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives.* **116**, 1111-1119.

Alan, J.I. (1991). Recent developments in non-parametric density estimation.*Journal of the American Statistical Association*, **86**, 205-224.

Anderson, N. and Titterington, D. (1997). Some methods of investigating spatial clustering with epidemiological applications. *Journal of the Royal Statistical Society Series A*, **169**, 87-105.

Baddeley, A., Turner, R., Moller, J. and Hazelton, M.L. (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society, Series B* **67**, 617-666.

Bartlett, M.S. (1963). Statistical estimation of density functions. *Sankhyā* Series A **25**, 245-254.

Berke, O. (2005). Exploratory spatial relative risk mapping. *Preventive Veterinary Medicine*, **71**, 173-182.

Bhattacharya, P. K. (1967). Estimation of a probability density function and its derivatives. *Sankhyā* Series A **29**, 373-382.

Bithel, J. F. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine*, **9**, 691-701.

Bithel, J. F. (1991). Estimation of relative risk functions. *Statistics in Medicine*, **10**, 1745-1751.

Bithel, J. F. (1992). Statistical methods for analyzing point-source exposures. In *Geographical and Environmental Epidemiology:Methods for Small Area Studies*(eds P.Elliott, J. Cuzick, D. English and R. Stern), pp. 221-230. Oxford: Oxford University Press.

Borla, A. and Protopopescu, C. (2010). Nonparametric estimation of the fractional derivative of a distribution function. Working Papers halshs-00536979, HAL.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353-60.

Bowman, A. W., Hall, P. and Titterington, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, **71**, 341-351.

Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, **18**, 179-89.

Chacón, J. E., Duong, T and Wand, M.P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, **21**, 807 - 840.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807-823.

Clark, A. B. & Lawson, A. B. (2004). An evaluation of non-parametric relative risk estimators for disease maps. *Computational Statistics & Data Analysis*, **47**, 63-78.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829-836.

Davies, T. M. and Hazelton, M. L. (2010). Adaptive kernel estimation of spatial relative risk. *Statistics in Medicine*, **29**, 2423-2437.

Devroye, L. and Gyöorfi, L. (1985). Non-parametric density estimation: *The L_1 View*. Wiley, New York.

Diggle, P.J. (1985b). A kernel method for smoothing point process data. *Applied Statistics*, **34**, 138-147.

Diggle, P. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society Series A*, **153**, 349-362.

Diggle, P. and Rowlingson, B. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Soc. Series A*, **157**, 433-440.

Duong, T. and Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15**, 17-30.

Duong, T. and Hazelton, M. L. (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, **32**, 485-506.

Elliott, P., Hills, M., Beresford, J., Kleinschmidt, I., Jolley, D., Pattenden, S. et al. (1992b). Incidence of cancer of the larynx and lung near incinerators of waste solvents in Great Britain. *The Lancet*, **339**, 854-858.

Fan, J. & Gijbels, I. (1992). Variable bandwidth local linear regression smoothers. *Annals of Statistics*, **20**, 2008-2036.

Fan, J. (1992a). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998-1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, **21**, 196-216.

Fan, J., Heckman, N., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, **90**, 141-150.

Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications. London: Chapmam & Hall.

Fan, J., Farmen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society Series B*, **60**, 591-608.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis. Non-parametric discrimination:consistency properties. *International Statistical Review*, **57**, 238-47.

Gatrell, A. C. (1990). On modelling spatial point patterns in epidemiology: cancer of the larynx in Lancashire. North West Regional Research Laboratory, Research Report No. 9.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320-328.

Gordis, L. (2000). *Epidemiology*. Saunders, USA.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

Gregg, N. M. (1941). Congenital cataract following German measles in the mother. *Transactions of the Opthalmologic Society of Australia*, **3**, 35-46.

Hädle, W., Marron, J. S. and Wand, M. P. (1990). Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society Series B*, **52**, 223-232.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, **11**, 1156-1174.

Hall, P. and Marron, J. S. (1987a). Extent to which least-squares cross-validation minimises integrated squared error in non-parametric density estimation. *Probability Theory and Related Fields*, **74**, 567-81.

Hall, P. and Marron, J. S. (1988). Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields*, **80**, 37-49.

Hall, P. & Marron, J.S. (1991a). Local minima in cross-validation functions. *Journal of Royal Statistical Society Series B*, **53**, 245-252.

Hall, P. and Marron, J. S. (1991b). Lower bounds for bandwidth selection in density estimation. *Probability Theory and Related Fields*, **90**, 149-73.

Härdle, W. (1990a). *Smoothing techniques with implementation in S.* Springer-Verlag, New York.

Härdle, W. (1990b). Applied Nonparametric Regression. *Cambridge University Press*, Cambridge.

Hastie, T.J. and Tibshirani, R.J. (1990). Generalized additive models. *Chapman and Hall*, London.

Hazelton, M. L. (2008). Letter to the Editor: Kernel estimation of risk surfaces without the need for edge correction. *Statistics in Medicine*, **27**, 2269-2272.

Hazelton, M. L. and Davies, T. M. (2009). Inference Based on Kernel Estimates of the Relative Risk Function in Geographical Epidemiology. *Biometrical Journal*, **51**, 98-109.

Jones, M. C. (1994). On kernel density derivative estimation. *Communication in Statistics Theory and Methods*, **23**, 2133-2139.

Keller, A. Z. and Terris, M. (1965). The association of alcohol and tobacco with cancer of the mouth and pharynx. *American Journal of Public Health*, **55**, 1578-85.

Kelsall, J. E. and Diggle, P. J. (1995a). Kernel estimation of relative risk. *Bernoulli*, **1**, 3-16.

Kelsall, J. E. and Diggle, P. J. (1995b). Non-parametric estimation of spatial variation in relative risk. *Statistics Medicine*, **14**, 2335-2342.

Kelsall, J. E. and Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Applied Statistics*, **47**, 559-573.

Kile, G.A., Packham, J.M. and Elliott, H.J. (1989). Myrtle wilt and its possible management in association with human disturbance of rainforest in Tasmania. *New Zealand Journal of Forestry Science*, **19**, 256-264.

Lawson,A. B., Biggeri, A. and Dreassi, E. (1999). Edge effects in disease mapping. *In Disease Mapping and Risk Assessment for Public Health*, eds. A. Lawson et al., Chichester: Wley, 83-96.

Lawson, A.B., and Zhou (2005). Spatial statistical modelling of disease outbreaks with particular reference to the UK foot and mouth (FMD) epidemic of 2001. *Preventive veterinary medicine*, **71**, 141-146.

Lo, C.P. and Yeung, A. K. W. (2007). Concepts and techniques of geographic information systems(2eds). Upper Saddle River (NJ): Prentice-Hall, Inc.

Marron, J.S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of Royal Statistical Society Series B*, **56**, 653-671.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman and Hall, London.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **10**, 186-90.

Packham, J. M.(1994). Studies on myrtle wilt. *PhD Thesis*, University of Tasmania.

Park, B.U. & Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *Journal of American Statistical Association*, **85**, 66-72.

Prince M.I., Chetwynd, A., Diggle, P.J., Jarner, M., Metcalf, J.V. and James, O.F.W. (2001). The Geographical Distribution of Primary Biliary Cirrhosis in a Well-Defined Cohort. *Hepatology*, **34**, 1083-1088.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-7.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1346-1370.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria 2008, URL be *http://www.R-project.org. ISBN 3-900051-07-0.*

Sabel et al. (2000). Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. *Social Science and Medicine*, **50**, 1121- 1137.

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association*, **89**, 807-817.

Schuster, E. F. (1969). Estimation of a probability function and its derivatives. *Annals of Mathematical Statistics*, **40**, 1187-1195.

Scott, D. W. and Factor, L. E.(1981). Monte carlo study of three data-based nonparametric probability density estimators. *Journal of American Statistical Association*, **76**, 9-15.

Scott, D. W. and Thompson, J. R. (1983). Probability density estimation in higher dimensions. In Gentle, J.E (ed.). *Computer science abd Statistics:Proceedings of the fifteenth symposium on the interface.* Amsterdam: North Holland, 173-179.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Signorini, D. F. and Jones, M. C. (2004). Kernel estimators for univariate binary regression. *Journal of the American Statistical Association*, **99**, 119-126.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, New York: Springer.

Singh, R. S. (1987). MISE of kernel estimates of a density and its derivatives. *Statistics and Probability Letters*, **5**, 153-159.

Staniswalis, J. G. (1989a). Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, **84**, 284-288.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of Royal Statistical Society B*, **36**, 111-147.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, **5**, 595-620.

Stone, C. J. (1980). Optimal rates of convergence for non-parametric estimators. *Annals of Statistics*, **8**, 1348-1360.

Stone, C. J. (1982). Optimal global rates of convergence of nonparametric regression. *Annals of Statistics*, **10**, 1040-1053.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, 1285-1297.

Stone, R. (1988). Investigations of excess environmental risks around putative source: Statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649-660.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of American Statistical Association*, **85**, 470-7.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82**, 559-568.

Truett, J., Cornfield, J. and Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*, **20**, 511-524.

Wakefield, J. C. and Elliott, P. (1999). Issues in the Statistical Analysis of Small Area Health Data. *Statistics in Medicine*, **18**, 2377-2399.

Wand, M. P. and Jones, M. C.(1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of American Statistical Association*, **88**, 520-8.

Wand, M. P. and Jones, M. C.(1995). *Kernel Smoothing.* London:Chapman and Hall.

Wand, M. P. (2008). Spatial epidemiology: Where have we come in 150 years? In: Sui DZ, ed. Geospatial technologies and homeland security: research frontiers and future challenges. Netherlands: Springer. pp 257-282.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Series A*, **26**, 101-116.

Weisberg, S. (1980). Applied Linear Regression. New York: Wiley.

Wheeler D. C. (2007). A Comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. *International Journal of Health Geographics 2007*; **6**(13). DOI: 10.1186/1476-072X-6-13.